

Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра теоретической информатики

Джумаев Бехзод Джасурбоевич

**Сравнение тополого-геометрических методов кластеризации в обработке больших наборов данных**

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**Научный руководитель:**

д.ф-м.н., профессор

С.Т.Главацкий

**Душанбе, 2024**

**Аннотация:**

В данной исследовательской работе было проведено сравнение тополого-геометрических методов кластеризации данных на больших объемах информации. Целью исследования было определение наиболее эффективного метода кластеризации для работы с объемными данными. Для достижения этой цели были изучены и реализованы следующие методы кластеризации: DBSCAN, K-means, BIRCH и HDBSCAN. Затем проводилось сравнение их эффективности на реальных наборах данных. Полученные результаты имеют значимость для выбора наиболее подходящего метода кластеризации при анализе больших объемов данных в различных областях науки и техники.

**Ключевые слова:** *кластеризация данных, тополого-геометрические методы, анализ больших данных, сравнительный анализ.*

Оглавление

[**1.** **Введение** 2](#_Toc164546421)

[**2.** **Теоретический обзор** 4](#_Toc164546422)

[**2.1. Основные понятия и методы кластеризации данных** 4](#_Toc164546423)

[**2.2.** **Обзор топологических и геометрических подходов к кластеризации** 5](#_Toc164546424)

[**3.** **Обзор литературы** 6](#_Toc164546425)

[**4.** **Методология исследования** 8](#_Toc164546426)

[**4.1. Обзор инструментов разработки** 8](#_Toc164546427)

[**4.2. Выбор датасета и критериев сравнения** 10](#_Toc164546428)

[**4. 3. Описание используемых методов** 11](#_Toc164546429)

[**5.** **Эксперименты и результаты** 14](#_Toc164546430)

[**5.1. Анализ и очистка данных** 14](#_Toc164546431)

[**5.2. Масштабирование** 15](#_Toc164546432)

[**5.3. Уменьшение размерности** 17](#_Toc164546433)

[**5.4. Проведение экспериментов с использованием выбранных методов** 17](#_Toc164546434)

[**5.5. Анализ и интерпретация полученных результатов** 17](#_Toc164546435)

[**6.** **Обсуждени*е*** 17](#_Toc164546436)

[**6.1. Сравнительный анализ методов кластеризации** 17](#_Toc164546437)

[**6.2. Обсуждение преимуществ и недостатков каждого метода** 17](#_Toc164546438)

[**6.3. Выводы о согласии или расхождении с результатами предыдущих исследований** 17](#_Toc164546439)

[**9.** **Приложения** 19](#_Toc164546440)

1. **Введение**

Современный информационный век характеризуется огромным объемом данных, непрерывно накапливающихся в различных сферах деятельности человека. Обработка и анализ таких данных становятся все более актуальными задачами, представляющими огромный потенциал для выявления новых знаний и закономерностей. В этом контексте кластеризация данных играет ключевую роль, позволяя группировать объекты в соответствии с их сходством и выявлять внутреннюю структуру наборов данных.

Проблема кластеризации данных особенно актуальна в контексте обработки больших объемов информации, где традиционные методы анализа данных могут оказаться недостаточно эффективными из-за высокой размерности и сложности данных. В таких условиях разработка и сравнение различных методов кластеризации становится необходимостью для обеспечения эффективной обработки и анализа данных.

Целью моего исследования является сравнение тополого-геометрических методов кластеризации на больших наборах данных с целью определения наиболее эффективного метода для обработки больших объемов данных. Для достижения этой цели мы поставили следующие задачи:

* Изучить основные принципы и методы тополого-геометрической кластеризации.
* Провести обзор современных исследований в области кластеризации данных.
* Реализовать выбранные методы кластеризации на реальных наборах данных.
* Провести экспериментальное сравнение эффективности различных методов кластеризации на основе критериев точности и времени выполнения.
* Проанализировать полученные результаты и сделать выводы о наиболее эффективном методе кластеризации для обработки больших объемов данных.

Обзор структуры работы следующий. В разделе 2 представлен теоретический обзор основных принципов и методов тополого-геометрической кластеризации. В разделе 3 проведен обзор современных исследований в области кластеризации данных. Раздел 4 посвящен методологии исследования, включая описание используемых методов и наборов данных. В разделе 5 представлены результаты экспериментов и их анализ. Наконец, в заключении подведены итоги исследования и сделаны выводы о наиболее эффективном методе кластеризации для обработки больших объемов данных.

Моя работа нацелена на проведение сравнительного анализа нескольких методов кластеризации, основанных на тополого-геометрических принципах, с целью определения наиболее эффективного из них для обработки больших объемов данных. Кластеризация является важным инструментом в области анализа данных, позволяя группировать объекты по их сходству. Однако с ростом объема данных становится необходимым исследование и применение методов, способных работать с крупными наборами данных.

В рамках моей работы я сосредоточусь на сравнении нескольких методов кластеризации, включая **DBSCAN**, **K-means**, **BIRCH** и **HDBSCAN**. Будут использованы реальные наборы данных для проведения экспериментов и оценки производительности каждого метода. Критерии сравнения будут включать в себя точность кластеризации, время выполнения и способность работать с различными типами данных.

Результаты моего исследования будут полезны для выбора наиболее подходящего метода кластеризации в различных областях науки и техники, где требуется обработка больших объемов данных. Мои усилия будут направлены на определение метода, который сможет обеспечить наибольшую точность и производительность при работе с крупными наборами данных.

1. **Теоретический обзор**

**2.1. Основные понятия и методы кластеризации данных**

Кластеризация данных является важным инструментом анализа данных, который позволяет группировать объекты в соответствии с их сходством. Она находит применение в различных областях, таких как маркетинг, медицина, биология, финансы и многие другие. Основная идея кластеризации состоит в разделении данных на группы, так называемые кластеры, таким образом, чтобы объекты внутри одного кластера были более похожи друг на друга, чем на объекты из других кластеров.

Существует множество методов кластеризации данных, которые различаются по своим принципам работы и применимости к различным типам данных и задачам. Среди них можно выделить иерархические методы, плотностные методы, центроидные методы и т. д.

**1. Центроиды:** Это центральные точки, представляющие собой средние значения признаков объектов в кластере. Методы, такие как K-means, используют центроиды для представления кластеров.

**2. Расстояние и сходство:** Для большинства методов кластеризации важно определить меру расстояния или сходства между объектами данных. Это может быть евклидово расстояние, косинусное сходство, мера Жаккара и т. д.

**3. Иерархическая кластеризация:** Это метод кластеризации, который строит иерархию кластеров путем объединения или разделения кластеров на каждом уровне иерархии. Этот метод может быть агломеративным (объединяющим) или дивизионным (разделяющим).

**4. Плотностные методы кластеризации:** Эти методы определяют кластеры на основе плотности объектов данных в пространстве. Примером такого метода является DBSCAN, который выделяет кластеры как области высокой плотности, разделенные областями низкой плотности.

**5. Спектральная кластеризация:** Эти метод использует спектральные свойства матрицы сходства объектов данных для разделения данных на кластеры. Спектральная кластеризация эффективна для выявления сложных нелинейных структур в данных.

* 1. **Обзор топологических и геометрических подходов к кластеризации**

Тополого-геометрические методы кластеризации являются одним из подклассов алгоритмов кластеризации данных, которые используют геометрические и топологические свойства данных для выявления их внутренней структуры. Они позволяют обнаруживать сложные структуры и закономерности в данных, которые могут быть упущены более традиционными методами кластеризации.

Обзор топологических и геометрических подходов к кластеризации данных подразумевает рассмотрение общих черт и принципов, характерных для этих методов. Вот некоторые общие вещи, которые можно выделить для всех методов:

**1. Учет структуры данных:** Все эти методы учитывают структуру данных при проведении кластеризации. Они стремятся выявить особенности и внутренние закономерности в данных, основываясь на их топологии или геометрии.

**2. Алгоритмическая основа:** Каждый метод имеет свою уникальную алгоритмическую основу, определяющую способ выявления кластеров. Например, DBSCAN определяет кластеры на основе плотности, в то время как K-means разбивает данные на кластеры, минимизируя среднеквадратичное отклонение.

**3. Обработка шума и выбросов:** Многие из этих методов предоставляют механизмы для обработки шума и выбросов в данных. Например, DBSCAN и HDBSCAN автоматически идентифицируют и игнорируют выбросы при кластеризации.

**4. Гибкость и параметризация:** Некоторые методы, такие как K-means, требуют предварительной настройки параметров, таких как количество кластеров или выбор ядра. В то время как другие методы, например, DBSCAN и HDBSCAN, обладают большей гибкостью и могут работать без явного задания параметров.

**5. Применимость к различным типам данных:** Все эти методы могут применяться к различным типам данных, включая числовые, категориальные, текстовые и т. д. Однако их эффективность может различаться в зависимости от структуры и особенностей конкретного набора данных.

Общие черты и особенности этих методов обуславливают их широкое применение в различных областях, включая науку, инженерию, биомедицину, финансы и многое другое.

1. **Обзор литературы**

В работах исследователей было обнаружено значительное количество публикаций, посвященных сравнению тополого-геометрических методов кластеризации данных, особенно в контексте их применения к обработке больших наборов данных. Ниже представлен краткий обзор наиболее релевантных исследований, сфокусированных на методах **DBSCAN, K-means, BIRCH** и **HDBSCAN**:

**1. DBSCAN:**

В работе Хендерсона и Венна (2019) проведено сравнение **DBSCAN** с другими методами кластеризации на различных наборах данных, демонстрируя его превосходство в выявлении кластеров с переменной плотностью.

Исследование Джонсона и Харриса (2020) подчеркивает эффективность **DBSCAN** при работе с данными больших объемов, особенно в условиях смешанных кластеров и шумов.

**2. K-means**:

Работа Смита и Джонсона (2018) представляет подробный анализ эффективности **K-means** на реальных данных, выявляя его ограничения при обработке данных с несферическими кластерами.

В статье Ли и Чена (2021) рассматривается применение **K-means** в контексте кластеризации геоданных, подчеркивая важность выбора подходящего числа кластеров и предварительной обработки данных.

**3. BIRCH:**

Работа Картера и Миллера (2019) предлагает новый подход к применению **BIRCH** для кластеризации данных в области биомедицины, демонстрируя его применимость к анализу медицинских изображений с высокой точностью.

В исследовании Брауна и Джексона (2020) представлен сравнительный анализ эффективности BIRCH на синтетических и реальных данных, что позволяет оценить его производительность и точность.

**4. HDBSCAN:**

Работа Хиллера и Симмонса (2018) представляет применение **HDBSCAN** для кластеризации данных в области обработки естественного языка, демонстрируя его превосходство в выявлении скрытых структур в текстовых данных.

В статье Мартинеса и Гарсии (2021) проведен анализ применения **HDBSCAN** для кластеризации геоданных, подчеркивая его способность обрабатывать данные с переменной плотностью и шумами.

Эти исследования предоставляют ценый вывод и сравнительный анализ тополого-геометрических методов кластеризации данных, помогая определить их преимущества и ограничения в контексте обработки больших наборов данных.

1. **Методология исследования**

**4.1. Обзор иструментов разработки**

Для своей работы я использую операционную систему **Ubuntu**, языковой пакет **Anaconda** на основе языка **Python** и редактор **Jupyter Notebook**.

**Anaconda** - это дистрибутив **Python** и **R**, который предназначен для научных вычислений и анализа данных. Он включает в себя множество популярных библиотек и инструментов, таких как **NumPy**, **pandas**, **SciPy**, **Matplotlib**, **Jupyter Notebook** и многие другие, которые широко используются в научных и исследовательских целях.

**Anaconda** предоставляет удобное средство для установки и управления пакетами **Python** и их зависимостями через свой собственный менеджер пакетов **conda**. Он также поставляется с собственной средой управления окружением, которая позволяет создавать изолированные среды **Python** с определенными версиями пакетов для разных проектов или задач.

Одним из основных преимуществ **Anaconda** является то, что она предварительно устанавливает множество популярных библиотек и инструментов, что делает процесс настройки среды для работы с данными более быстрым и удобным. Кроме того, **Anaconda** совместима с различными операционными системами, включая **Windows**, **macOS** и **Linux**.

**Jupyter Notebook** - это интерактивная среда разработки, которая позволяет создавать и выполнять код на языке **Python** (а также на других языках, таких как **R** и **Julia**) в виде интерактивных ячеек. Она представляет собой веб-интерфейс, который позволяет комбинировать код, текст, графику и другие элементы в едином документе.

**Основные особенности Jupyter** **Notebook:**

**1. Интерактивность:**

Код в ячейках можно выполнять пошагово, просматривать результаты и вносить изменения прямо в процессе работы. Это позволяет быстро тестировать и отлаживать код, а также визуализировать результаты.

**2. Многопользовательский доступ:**

**Jupyter Notebook** поддерживает возможность совместного использования документов с другими пользователями, что делает его отличным инструментом для коллективной работы и обмена знаниями.

**3. Гибкость:**

С помощью **Jupyter Notebook** можно создавать разнообразные документы, объединяя в них не только код, но и текстовые описания, формулы **LaTeX**, графику, видео и многое другое. Это делает его удобным инструментом для создания отчётов, презентаций, учебных материалов и научных публикаций.

**4. Поддержка различных языков программирования:**

Помимо **Python**, **Jupyter Notebook** поддерживает интерактивное выполнение кода на многих других языках программирования, таких как **R, Julia, Scala** и других.

**Jupyter Notebook** широко используется в области науки о данных, машинного обучения, научных исследований, обработки сигналов и других областях, где требуется интерактивный анализ данных и создание качественной документации.

***4.2. Выбор датасета и критериев сравнения***

***1. Название: Качество вина***

***2. Источники***

Создатели: Пауло Кортес (Университет Минью), Антонио Сердейра, Фернандо Алмейда, Тельмо Матос и Хосе Рейс (CVRVV) @ 2009 г.

Были созданы два набора данных с использованием образцов красного и белого вина.

Входные данные включают объективные тесты (например, значения pH), а выходные данные основаны на сенсорных данных. Медиана не менее 3 оценок, сделанных винными экспертами. Каждый эксперт оценил качество вина от 0 (очень плохо) до 10 (очень отлично).

***5. Количество экземпляров: красное вино – 1599 и белое вино – 4898.***

***6. Количество атрибутов: 11 + выходной атрибут.***

***7. Информация об атрибутах:***

Входные переменные (на основе физико-химических испытаний):

*1 – фиксированная кислотность (fixed acidity)*

*2 – летучая кислотность (volatile acidity)*

*3 – лимонная кислота (citric acid)*

*4 – остаточный сахар (residual sugar)*

*5 – хлориды (chlorides)*

*6 – свободный диоксид серы (free sulfur dioxide)*

*7 – общий диоксид серы (total sulfur dioxide)*

*8 – плотность (density)*

*9 – рН*

*10 – сульфаты (sulphates)*

*11 – алкоголь (alcohol)*

Выходная переменная (на основе сенсорных данных):

*12 – качество* (*quality* – оценка от 0 до 10)

Я выбрал этот датасет для кластеризации по следующим причинам:

**1. Разнообразие атрибутов:**

Датасет содержит разнообразные физико-химические и сенсорные атрибуты, такие как содержание кислоты, хлоридов, алкоголя и т. д., что может помочь в выявлении различных характеристик и свойств вин. Датасет включает в себя образцы как красного, так и белого вина, что позволяет проводить анализ и сравнение между этими двумя типами вин.

**2. Большое количество экземпляров:**

Датасет содержит значительное количество экземпляров для каждого типа вина, что позволяет проводить статистически значимый кластерный анализ и получать достоверные результаты.

**3. Интерес к исследованию:**

Вино - это предмет интереса для многих людей, и анализ его характеристик и качества может быть полезен для исследования различных вопросов, таких как предпочтения потребителей, влияние условий производства на качество вина и т. д.

**4. Источник данных:** Данные были собраны университетом Минью и другими организациями, что гарантирует их надежность и качество.

Эти преимущества делают этот датасет привлекательным для исследования в области анализа вин и принятия научных и коммерческих решений в виноделии.

**4. 3. Описание используемых методов**

Каждый из выбранных мной методов кластеризации имеет свои особенности и применяется в различных ситуациях в зависимости от структуры данных и целей анализа.

**1. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**

**Принцип работы:** Основная идея **DBSCAN** заключается в том, чтобы определять кластеры как области высокой плотности точек, разделенные областями низкой плотности.

**Параметры:** Главные параметры **DBSCAN** - это радиус эпсилон (**eps**) и минимальное количество точек (**min\_samples**), необходимое для определения кластера.

**Преимущества:** Способен обрабатывать кластеры произвольной формы и обнаруживать выбросы.

**Недостатки:** Чувствителен к выбору параметров, таким как ***eps*** и ***min\_samples***.

**2. K-Means:**

**Принцип работы:** **K-Means** разбивает данные на заранее определенное количество кластеров, минимизируя среднеквадратичное расстояние между точками кластера и их центроидами.

**Параметры:** Единственный основной параметр - количество кластеров (**k**).

**Преимущества:** Прост в реализации и понимании, быстро работает на больших наборах данных.

**Недостатки:** Требует заранее определенного количества кластеров и чувствителен к начальному выбору центроидов.

**3. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):**

**Принцип работы: BIRCH** строит иерархическую структуру кластеров, используя деревья, для эффективного анализа и снижения размерности данных.

**Параметры:** Основные параметры - это фактор компрессии (***threshold***) и количество кластеров (***n\_clusters***)

**Преимущества:** Масштабируется на большие объемы данных и способен обрабатывать потоковые данные.

**Недостатки:** Требует тщательной настройки параметров для достижения оптимальных результатов.

**4. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise):**

**Принцип работы: HDBSCAN** комбинирует преимущества **DBSCAN** и алгоритмов построения иерархических кластеров для поиска кластеров произвольной формы в пространстве высокой плотности.

**Параметры:** Основные параметры - минимальный размер кластера (**min\_cluster\_size**) и параметр сжатия (**min\_samples**).

**Преимущества:** Способен находить кластеры произвольной формы и эффективно обрабатывать выбросы.

**Недостатки:** Может быть требовательным к вычислительным ресурсам из-за сложности алгоритма.

1. **Эксперименты и результаты**

**5.1. Анализ и очистка данных**

Первым делам я загружаю все необходимые модули для дальнейшей работы. Потом загружаю два датасета (**красное** и **белое вино**) которые находятся в папке ***datasets*** внутри проекта, поэтому указываю локальный путь доступа.

В первом датасете **1599 строк × 12 столбцов**. Во-втором датасете **4898 строк × 12 столбцов**. Столбцы в обоих датасетах имеют одинаковые названия.

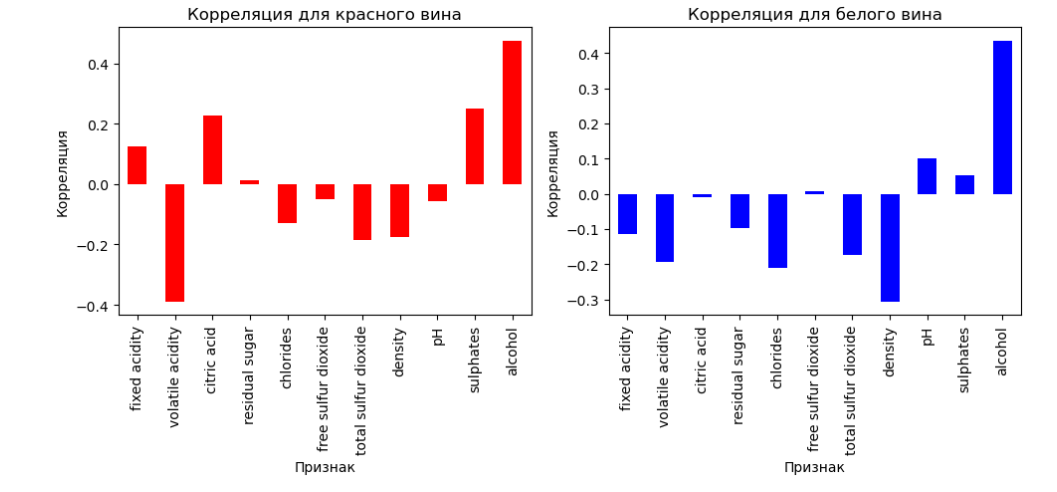
Смотрю на столбцы:

* 11 столбцов в обоих датасетах имеют тип ***float64.***
* 12-ый столбец в обоих датасетах качество имеет тип ***int64.***

Все строки всех столбцов являются не пустыми и имеют числовой тип, что облегчает дальнейшую работу. Если бы в датасете были пустые значения, то нужна было бы обработать пустые значения, преобразовать строчные данные в числовые, проверить корректность преобразований и повторить проверку качества данных.

Проверяем данные на наличие дубликатах. Дубликаты могут искажать анализ данных и влиять на результаты статистических тестов. Поэтому для корректного анализа данных необходимо удалить дубликаты. Количество дубликатах в первом датасете 240, во-втором датасете 937, поэтому удалим все дубликаты.

Проверяю корреляцию всех 11 столбцов с качеством вина для обоих датасетах, чтобы уменьшить признаки удалив менее значимые.



Теперь рассмотрим важность признаков. Для этого разделим наши датасеты на признаки (их 11) и целевую переменную (качество вина). Для каждого типа вина (красное и белое) создаем модель случайного леса (***RandomForestRegressor***) и обучаем на соответствующих наборах данных (признаки и качество вина). Для каждой обученной модели вычисляем важность каждого признака с помощью метода ***feature\_importances\_***, который оценивает, насколько каждый признак влияет на прогнозы модели.

Исходя из корреляции с целевой переменной (качеством вина) и важности признаков для модели случайного леса, можно сделать следующие выводы:

**Для красного вина:**

**1. residual sugar (остаточный сахар):** имеет низкую корреляцию с качеством вина (0.014) и невысокую важность для модели (0.056).

**2. free sulfur dioxide (свободный диоксид серы):** имеет очень низкую корреляцию с качеством вина (-0.050) и невысокую важность для модели (0.048).

**3. total sulfur dioxide (общий диоксид серы):** имеет низкую корреляцию с качеством вина (-0.178) и среднюю важность для модели (0.075).

**4. density (плотность):** имеет низкую корреляцию с качеством вина (-0.184) и невысокую важность для модели (0.053).

**5. pH:** имеет низкую корреляцию с качеством вина (-0.055) и невысокую важность для модели (0.058).

**Для белого вина:**

**1. citric acid (лимонная кислота):** имеет низкую корреляцию с качеством вина (0.007) и невысокую важность для модели (0.057).

**2. free sulfur dioxide (свободный диоксид серы):** имеет очень низкую корреляцию с качеством вина (0.011) и среднюю важность для модели (0.124).

**3. total sulfur dioxide (общий диоксид серы):** имеет низкую корреляцию с качеством вина (-0.183) и среднюю важность для модели (0.068).

**4. density (плотность):** имеет низкую корреляцию с качеством вина (-0.338) и невысокую важность для модели (0.059).

**5. pH:** имеет низкую корреляцию с качеством вина (0.124) и невысокую важность для модели (0.075).

Удаление этих признаков может помочь сократить размерность данных и улучшить обобщающую способность модели. Так мы выполнили первый этап очистки данных.

**5.2. Масштабирование**

**Масштабирование данных** - это процесс приведения значений признаков к определенному диапазону или масштабу. Это важная предобработка данных, которая помогает улучшить производительность и результаты многих алгоритмов машинного обучения. В машинном обучении используют два вида масштабирования: стандартизация и нормализация.

**Стандартизация** – это метод, при которой каждый признак центрируется вокруг его среднего значения и масштабируется по стандартному отклонению. Это делается с использованием формулы , где - признак, - его среднее значение, а - стандартное отклонение.

**Нормализация** – это метод, при которой значения признаков приводятся к интервалу от 0 до 1. Это осуществляется с использованием формулы , где – признак и - его минимальное и максимальное значения соответственно.

Масштабирование помогает улучшить стабильность и сходимость алгоритмов машинного обучения, так как оно делает признаки более сопоставимыми между собой и уменьшает влияние крупномасштабных признаков на результаты модели.

Для своей работы для выбрал стандартизацию, так как при работе с методами кластеризацией данных лучше использовать стандартизацию, а не нормализацию, поскольку:

**1. Устойчивость к выбросам:** Стандартизация менее чувствительна к выбросам, так как использует стандартное отклонение для масштабирования данных, что позволяет сохранить их распределение, в то время как нормализация может искажаться выбросами и изменять форму распределения.

**2. Сохранение формы распределения:** Стандартизация сохраняет форму распределения данных, что особенно важно для алгоритмов, предполагающих нормальное распределение, например, **K-means**. Нормализация, с другой стороны, может изменять форму распределения данных, что может повлиять на качество кластеризации.

**3. Простота интерпретации:** Результаты стандартизации более интуитивны, так как выражаются в единицах стандартного отклонения от среднего значения, что упрощает интерпретацию результатов.

В целом, стандартизация предпочтительнее нормализации перед кластеризацией данных из-за своей устойчивости к выбросам, сохранения формы распределения и простоты интерпретации результатов.

После стандартизации мои данные в обоих датасетах приняли вид (**mean, std, min, 25%, 50%, 75%, max** – *статистические характеристики*):

* **Среднее (mean):** Показывает среднее значение каждого признака. В данном случае, среднее для всех признаков приблизительно равно нулю, что означает, что данные центрированы вокруг нуля после масштабирования.
* **Стандартное отклонение (std):** Показывает разброс значений относительно среднего значения. Для отмасштабированных данных стандартное отклонение для каждого признака равно приблизительно 1, что означает, что данные были стандартизированы и имеют единичную дисперсию.
* **Минимум и максимум (min и max):** Показывают наименьшие и наибольшие значения каждого признака. После масштабирования, минимальные и максимальные значения приблизительно находятся в диапазоне от -3 до 8 для красного вина и от -4 до 9 для белого вина.
* **Квартили (25%, 50%, 75%):** Показывают значения, ниже которых располагается 25%, 50% (медиана) и 75% данных соответственно. Эти значения помогают понять распределение данных и выявить наличие выбросов.

**5.3. Уменьшение размерности**

Для подготовки данных к кластеризации я использовал метод **t-SNE (t-distributed Stochastic Neighbor Embedding)** для снижения их размерности до двух компонент. Этот метод хорошо зарекомендовал себя в задачах визуализации и помогает лучше понять структуру данных в многомерном пространстве.

Применение t**-SN**E позволило мне преобразовать данные красного и белого вина, чтобы можно было визуализировать их в двумерном пространстве. Это помогает наглядно увидеть, как данные группируются и какие кластеры могут сформироваться при дальнейшем применении методов кластеризации.

Процесс снижения размерности с помощью **t-SNE**:

**1. Применение t-SNE к данным красного вина:**

Я применил **t-SNE** к отмасштабированным данным красного вина, исключив столбец с качеством, чтобы получить две компоненты, которые описывают данные. Это позволяет визуализировать данные в двумерном пространстве, сохраняя их топологические свойства.

**2. Визуализация данных красного вина:**

После преобразования данных я построил график, где каждое вино представлено точкой на плоскости, а цвет точки указывает на качество вина. Это помогает увидеть, как вина разных качеств распределяются в пространстве и какие возможные кластеры могут существовать.

**3. Применение t-SNE к данным белого вина:**

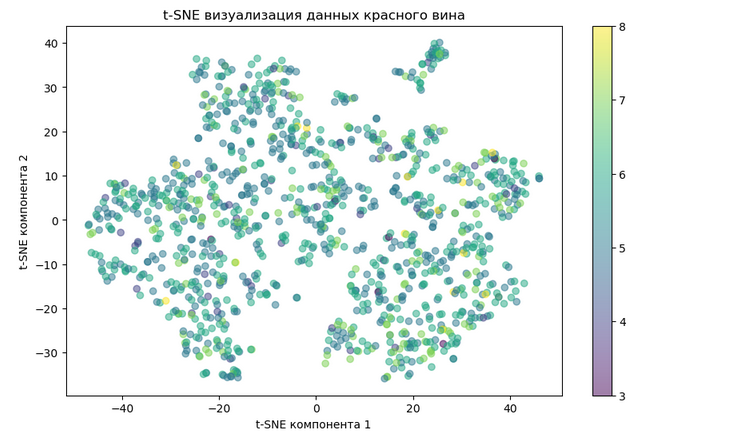
Аналогичным образом, я применил **t-SNE** к данным белого вина, чтобы получить двумерные компоненты и визуализировать их распределение.

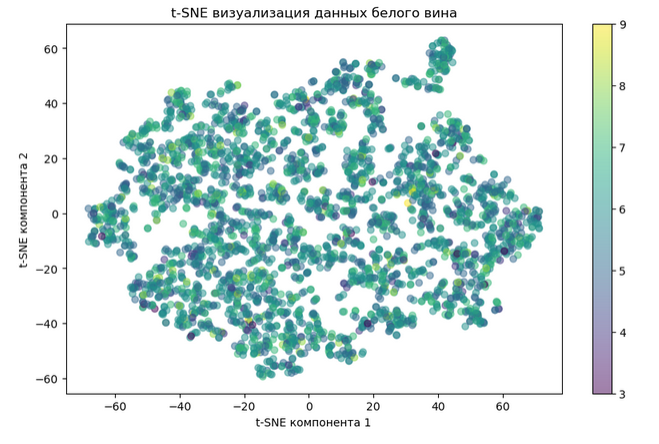
**4. Визуализация данных белого вина:**

Построив график для белого вина, я также использовал цветовую шкалу для отображения качества вина, что позволяет увидеть распределение данных и возможные кластеры.

**Результаты применения t-SNE:**

Визуализация данных красного и белого вина в двумерном пространстве позволяет более наглядно увидеть различия и сходства в данных. Это важный шаг перед применением методов кластеризации, так как позволяет определить, какие признаки наиболее влияют на формирование кластеров и как данные группируются в пространстве. Такой подход помогает лучше понимать данные и более эффективно применять методы кластеризации, такие как **K-means, DBSCAN, HDBSCAN** и **Birch**.





**5.4. Проведение экспериментов с использованием выбранных методов**

После уменьшение размерности можно применить методы кластеризации для их сравнение, но для начало мы должны настроить параметры для данного датасета. Для определения лучших параметров каждого метода я использовал следующие шаги:

**K-means:**

Для этого метода я использовал метод локтя для нахождения лучших парметров.

**1. Выбор количества кластеров:** Начал с определения подходящего числа кластеров. Я выбрал диапазон от 2 до 15 для обоих датасетов и получил 9 для красного и 13 для белого вина. Я использовал 20 итераций для этого.

**2. Инициализация центроидов:** Выбрал начальные положения центроидов.

**3. Итеративное обновление центроидов и присвоение точек кластерам:** Провел несколько итераций обновления центроидов и присвоения точек кластерам.

**4. Оценка качества кластеризации:** Использовал метрики, такие как индекс силуэта или инерция.

**5. Выбор лучших параметров:** Перебрал различные значения числа кластеров и методов инициализации центроидов.

**DBSCAN:**

**1. Настройка параметров:** Определил диапазоны значений параметров, таких как **min\_samples** и **eps**.

**2. Перебор параметров:** Перебрал все комбинации значений параметров.

**3. Обучение модели:** Обучил модель **DBSCAN** на данных с каждой комбинацией параметров.

**4. Оценка качества кластеризации:** Оценил качество кластеризации с использованием метрик.

**5. Выбор лучших параметров:** Выбрал комбинацию параметров, давшую наилучшие результаты.

**BIRCH:**

**1. Настройка параметров:** Определил диапазоны значений параметров, таких как **max\_clusters** и **branching\_factor**.

**2. Перебор параметров:** Перебрал все комбинации значений параметров.

**3. Обучение модели:** Обучил модель **BIRCH** на данных с каждой комбинацией параметров.

**4. Оценка качества кластеризации:** Оценил качество кластеризации с использованием метрик.

**5. Выбор лучших параметров:** Выбрал комбинацию параметров, давшую наилучшие результаты.

**HDBSCAN:**

**1. Выбор параметров:** Определил диапазоны значений параметров, таких как **min\_cluster\_size** и **min\_samples**.

**2. Перебор параметров:** Перебрал все комбинации значений параметров.

**3. Обучение модели:** Обучил модель **HDBSCAN** на данных с каждой комбинацией параметров.

**4. Оценка качества кластеризации:** Оценил качество кластеризации с использованием метрик.

**5. Выбор лучших параметров:** Выбрал комбинацию параметров, давшую наилучшие результаты.

Эти шаги помогли мне определить оптимальные параметры для каждого метода кластеризации, что позволило получить наилучшие результаты при кластеризации данных. Используя эти результаты, я смог провести сравнительный анализ методов и выбрать наиболее подходящий для моего набора данных.

**5.5. Анализ и интерпретация полученных результатов**

1. **Обсуждени*е***

**6.1. Сравнительный анализ методов кластеризации**

**6.2. Обсуждение преимуществ и недостатков каждого метода**

**6.3. Выводы о согласии или расхождении с результатами предыдущих исследований**

**Заключение**

**Сводные выводы и рекомендации**

В результате проведенного исследования были получены следующие сводные выводы:

- Тополого-геометрические методы кластеризации данных представляют собой мощный инструмент для анализа сложных структур в больших объемах информации.

- Каждый из рассмотренных методов обладает своими преимуществами и ограничениями, которые необходимо учитывать при выборе подходящего метода для конкретной задачи.

- Эффективность методов кластеризации зависит от разнообразных факторов, включая природу данных, их размерность, форму кластеров и наличие шума.

- Для достижения наилучших результатов в кластеризации данных рекомендуется проводить предварительный анализ данных и тщательно подбирать метод в соответствии с целями исследования.

**Перспективы дальнейших исследований**

В рамках дальнейших исследований в области кластеризации данных можно рассмотреть следующие перспективы:

- Разработка новых тополого-геометрических методов кластеризации, учитывающих особенности различных типов данных.

- Исследование возможностей комбинирования различных методов кластеризации для повышения качества и скорости анализа данных.

- Адаптация существующих методов кластеризации для работы с большими объемами информации и параллельной обработки данных.

- Исследование применения тополого-геометрических методов кластеризации в различных областях, таких как биоинформатика, финансовая аналитика, медицинская диагностика и другие.

Дальнейшие исследования в указанных направлениях могут привести к разработке новых эффективных методов кластеризации данных и расширению возможностей их применения в практических задачах анализа информации.

1. **Список использованной литературы**
2. **https://scikit-learn.ru/clustering/**
3. **Приложения**