УДК 004.93

Репина С. И.,
ведущий аналитик данных
Skyeng
Российская Федерация, г.Москва
Repina S.I.,
Senior Data Analyst
Skyeng
Russia, Moscow

ПРОВЕРКА КАЧЕСТВА КЛАСТЕРОВ С ПОМОЩЬЮ СИЛУЭТНОГО АНАЛИЗА
VERIFICATION THE QUALITY OF CLUSTERS USING SILHOUETTE ANALYSIS

В данной статье рассматривается использование силуэтного анализа для оценки качества кластеров, полученных с помощью двух методов кластеризации: K-modes и иерархической кластеризации. Работа включает в себя сравнение результатов различных методов кластеризации и выбор наиболее подходящего из них на основе силуэтного анализа. Проведенный анализ демонстрирует, что иерархическая кластеризация обеспечивает более высокое качество сегментации данных по сравнению с методом K-modes, особенно при большом количестве кластеров. В статье также рассматриваются практические рекомендации по применению данных методов и силуэтного анализа для задач кластеризации в областях, таких как маркетинг и социальные науки.

This paper discusses the use of silhouette analysis to evaluate the quality of clusters obtained by two clustering methods: K-modes and hierarchical clustering. The work includes comparing the results of different clustering methods and

selecting the most appropriate one based on silhouette analysis. The analysis demonstrates that hierarchical clustering provides better quality of data segmentation compared to the K-modes method, especially when the number of clusters is large. The paper also discusses practical recommendations for applying these methods and silhouette analysis to clustering problems in areas such as marketing and social sciences.

Ключевые слова. *Кластеризация, K-modes, иерархическая кластеризация, силуэтный анализ, оценка качества кластеров, машинное обучение, анализ данных, сегментация данных, визуализация кластеров, большие данные.*

Keywords. *Clustering, K-modes, hierarchical clustering, silhouette analysis, cluster quality assessment, machine learning, data analysis, data segmentation, cluster visualization, big data.*

Введение (Introduction)

1) Актуальность выбора темы (Relevance of the topic selection)

В условиях возрастающего объема и сложности данных, методы кластеризации становятся все более важными инструментами для их анализа и обработки. Кластеризация позволяет автоматически группировать объекты на основе их характеристик, что особенно ценно для выявления скрытых структур и закономерностей в данных. Однако, несмотря на широкое распространение методов кластеризации, оценка качества формируемых кластеров остается сложной задачей. Силуэтный анализ выступает одним из ключевых методов, позволяющих количественно оценить качество кластеров и степень их внутренней согласованности.

2) Цель исследования (Purpose of the study)

Основная цель данного исследования заключается в анализе и сравнении результатов различных методов кластеризации с применением силуэтного анализа. В рамках исследования поставлены следующие задачи:

- Применение различных методов кластеризации, таких как K-modes и иерархическая кластеризация, к заданному набору данных.
- Расчет индексов силуэта для каждого метода и визуализация полученных результатов.
- Сравнительный анализ качества кластеров, полученных различными методами, на основе силуэтного анализа.
- Определение наиболее подходящего метода кластеризации для конкретного набора данных.

3) Значимость и актуальность исследования (Significance and relevance of the study)

Актуальность исследования связана с необходимостью повышения точности и надежности методов кластеризации в различных областях, таких как маркетинг, социальные науки и другие. В условиях работы с большими данными правильная сегментация и анализ кластеров позволяют более эффективно использовать ресурсы, улучшать качество принятия решений и разрабатывать более точные модели прогнозирования.

Силуэтный анализ предоставляет исследователям и практикам мощный инструмент для оценки качества кластеров, что способствует улучшению интерпретации и использования результатов кластеризации. Настоящее исследование вносит вклад в развитие методов оценки кластеров, предоставляя сравнительный анализ и рекомендации по выбору наиболее эффективного метода кластеризации в зависимости от особенностей данных.

Таким образом, работа представляет значительный интерес для специалистов, работающих с большими данными, и может быть полезна широкому кругу пользователей, стремящихся улучшить качество анализа и интерпретации данных в своих исследованиях и практических приложениях.

Обзор литературы (Literature review)

1) Методы кластеризации данных (Data clustering methods)

Кластеризация является одним из ключевых методов анализа данных, которая используется для группировки объектов на основе их схожести. Существует множество методов кластеризации, каждый из которых имеет свои преимущества и недостатки. Наиболее популярными являются методы K-means и иерархическая кластеризация.

K-means – один из самых широко используемых и простых методов кластеризации. K-means итеративно минимизирует сумму квадратов расстояний между объектами и центроидами кластеров. Основными недостатками метода являются необходимость задания количества кластеров заранее и чувствительность к выбросам и начальным условиям.

Иерархическая кластеризация – метод, не требующий предварительного задания количества кластеров и создающий вложенные кластеры, представленные в виде дендрограммы. Основными недостатками метода являются его вычислительная сложность и чувствительность к шуму в данных.

2) Критерии оценки качества кластеров (Criteria for assessing the quality of clusters)

Эффективность кластеризации можно оценить различными способами, такими как внутрикластерная дисперсия, межкластерное расстояние, коэффициент Джини, индекс Дэвиса-Булдина и другие. Однако, одним из наиболее универсальных является силуэтный анализ.

Индекс силуэта показывает, насколько хорошо каждый объект соответствует своему кластеру по сравнению с другими кластерами. Индекс силуэта имеет значения в диапазоне от -1 до 1, где высокие значения указывают на хорошую кластеризацию, а низкие или отрицательные значения указывают на возможные проблемы с кластеризацией.

3) Силуэтный анализ: концепция и применение (Silhouette analysis: concept and application)

Силуэтный анализ был предложен Питером Й. Русселя (Peter J. Rousseeuw) в 1987 году и с тех пор широко используется для оценки качества кластеризации.[1] Индекс силуэта для каждого объекта вычисляется на основе среднего внутрикластерного расстояния и среднего межкластерного расстояния до ближайшего кластера. Среднее значение индекса силуэта для всех объектов в наборе данных используется для оценки общего качества кластеризации.

Силуэтный анализ используется в различных областях, включая маркетинг, социальные науки и машинное обучение. Он помогает исследователям и практикам оценить, насколько хорошо данные сегментированы, и выбрать наиболее подходящий метод кластеризации для конкретной задачи.

4) Существующие подходы и исследования в области оценки кластеров (Existing approaches and studies on cluster assessment)

Существует множество исследований, посвященных разработке и применению методов оценки кластеров. В литературе можно найти сравнительные анализы различных методов, а также рекомендации по выбору подходящего метода для конкретных задач.

Многие исследования сравнивают силуэтный анализ с другими методами, такими как индекс Дэвиса-Булдина, коэффициент Джини и внутрикластерная дисперсия.[2] Эти исследования показывают, что силуэтный анализ является одним из наиболее информативных и универсальных методов для оценки качества кластеров.

В литературе существует множество примеров применения силуэтного анализа в различных областях. Например, в маркетинге силуэтный анализ используется для сегментации клиентов и оценки эффективности маркетинговых стратегий.

[1] Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987 20: 53-65

[2] Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review. ACM Computing Surveys 1999 31(3): 264-323

Методы (Methods)

1) Описание данных (Data description)

1.1. Источник данных (Data Source)

В данном исследовании были использованы данные из базы данных пользователей, желающих приобрести курсы онлайн-школы, представляющие собой id юзера, тип устройства, с которого он зашел, регион, ответы на вопросы №1,3,4,5 опроса сразу после подачи заявки. Данные охватывают период с 19.04.2024 по 12.06.2024. Объем данных составляет 3102 записей, каждая из которых содержит 6 признаков.

1.2. Описание признаков и их предварительная обработка (Description of features and their preprocessing)

Данные включают следующие ключевые признаки:

1. **Признак 1**: Девайс пользователя
2. **Признак 2**: Регион пользователя
3. **Признак 3**: Ответ на вопрос №1
4. **Признак 4**: Ответ на вопрос №3
5. **Признак 5**: Ответ на вопрос №4
6. **Признак 6**: Ответ на вопрос №5

Признаки были предварительно обработаны для обеспечения корректного выполнения методов кластеризации. Основные этапы предварительной обработки включали:

1. **Очистка данных**: Удаление или замена пропущенных значений, обработка выбросов.
2. **Кодирование категориальных признаков**:

№ (вес признака)	**device_type**	**region**	**q1 Родитель или ребенок?**	**q3 Есть опыт онлайн?**	**q4 Как срочно?**	**q5 Еще что-то посещаете?**
0	other	null, Киргизия, Молдавия, Туркмения, Узбекистан, Таджикистан	отвал	отвал	отвал	отвал

1	мобилка	Регионы РФ + другие страны	1й ответ Ребенок	1й ответ Нет	1й ответ Не торопимся	1й ответ Нет
2	десктоп	Москва и область	2й ответ Родитель	2й ответ Да	2й ответ Неделя-две	2й ответ Да
3					3й ответ Скорее	

Рис. 1. Таблица: Разметка признаков

Fig. 1. Table: Markup of features

3. **Проверка на корреляцию признаков**: Расчет коэффициента корреляции Пирсона для каждой возможной пары признаков.

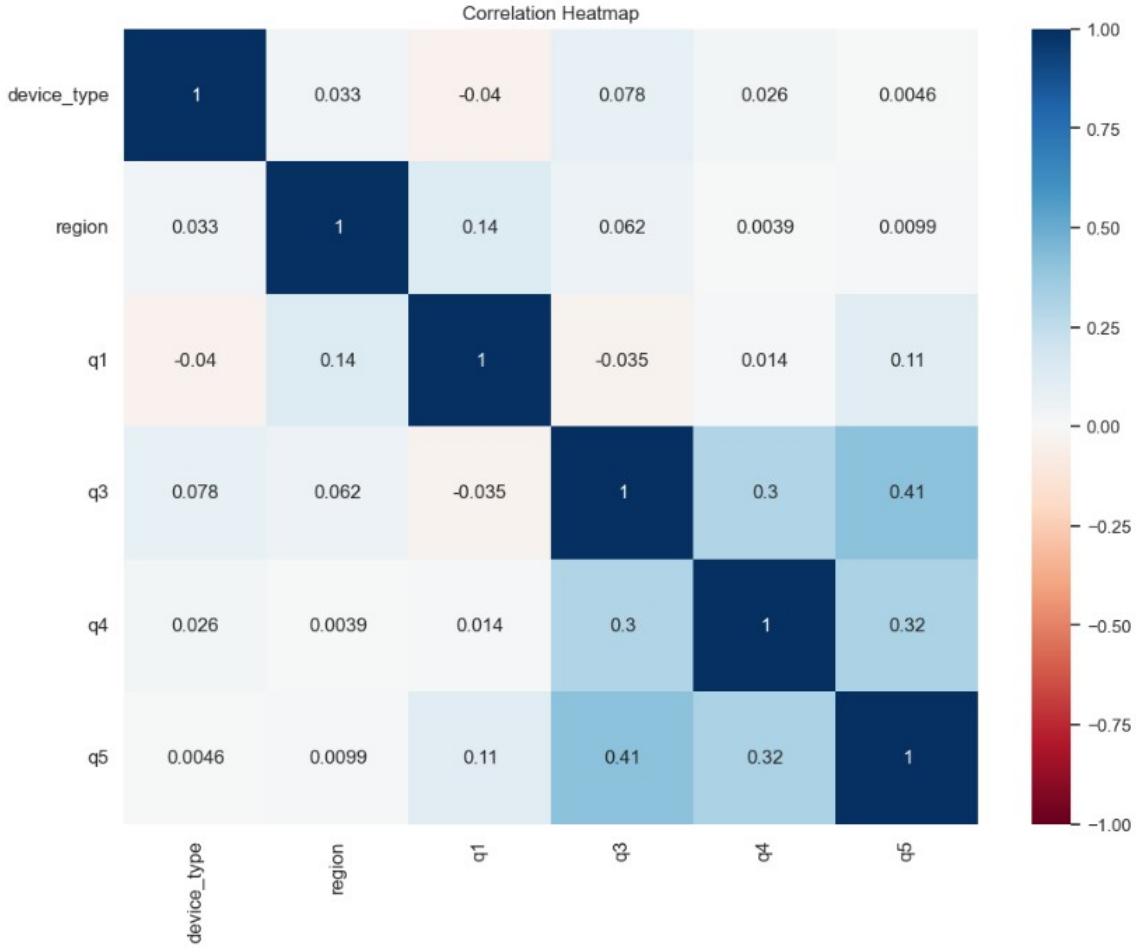

Рис. 2. График: Тепловая карта корреляций признаков

Fig. 2. Graph: Heat map of trait correlations

На графике видно, что разброс коэффициентов корреляции лежит в диапазоне от -0.04 до 0.41, то есть ближе к 0. Это говорит о том, что

признаки плохо коррелируют между собой. Значит можно продолжать исследование, используя все имеющиеся признаки.

2) Методы кластеризации (Clustering methods)

2.1. Описание выбранных методов (Description of selected methods)

Для проведения кластеризации данных были выбраны следующие методы:

K-modes

Метод K-modes является расширением метода K-means, предназначенным для работы с категориальными данными. Вместо вычисления среднего значения (mean) для кластеров, метод K-modes использует моду (mode), что позволяет эффективно обрабатывать категориальные признаки.[1]

Иерархическая кластеризация

Иерархическая кластеризация представляет из себя древовидную структуру (дендрограмму), изображающую вложенные кластеры.

2.2. Параметры и настройки моделей (Parameters and model settings)

Для оптимальной работы методов кластеризации были выбраны следующие параметры и настройки:

K-modes:

- Определение количества кластеров: Использование метода локтя (Elbow Method) и силуэтного анализа.
- Инициализация центров кластеров: Методом Хуанга (Huang) или случайной инициализацией.

Иерархическая кластеризация:

- Метод объединения кластеров: Метод Уорда (Ward), минимизирующий сумму квадратов расстояний внутри кластеров.

[1] K-means and K-modes: A Comparison. (n.d.). Retrieved from https://www.datascience.com/blog/k-means-and-k-modes-clustering

- Метрика расстояния: Евклидово расстояние для расчета расстояний между объектами.
- Определение количества кластеров: Использование дендрограммы и метода разреза (cut-off) на уровне, где расстояние между кластерами максимально увеличивается.

3) Применение силуэтного анализа (Application of silhouette analysis)

3.1. Теоретические основы силуэтного анализа (Theoretical basis of silhouette analysis)

Силуэтный анализ был введен Питером Й. Русселя (Peter J. Rousseeuw) в 1987 году и стал важным инструментом для оценки качества кластеризации. Индекс силуэта для каждого объекта измеряет, насколько хорошо этот объект соответствует своему кластеру по сравнению с другими кластерами. Значение индекса силуэта варьируется от -1 до 1:

- Значение близкое к 1 указывает на то, что объект находится внутри своего кластера и далеко от ближайшего кластера.
- Значение близкое к 0 указывает на то, что объект находится на границе между двумя кластерами.
- Отрицательное значение указывает на то, что объект, вероятно, присвоен неправильному кластеру.

Для каждого объекта i индекс силуэта s(i) вычисляется по формуле:

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

где:

- a(i) — это среднее расстояние между объектом i и всеми другими объектами его кластера.
- b(i) — это минимальное среднее расстояние от объекта i до объектов других кластеров (то есть до ближайшего кластера).

Средний индекс силуэта по всем объектам используется для оценки общего качества кластеризации. Значение близкое к 1 указывает на хорошо выполненную кластеризацию.

3.2. Процесс расчета индексов силуэта (The process of calculating silhouette indices)

Процесс расчета индексов силуэта включает несколько этапов:

1. **Выбор метрики расстояния**:
   - Для различных типов данных могут использоваться различные метрики расстояния.
2. **Расчет внутрикластерного расстояния (a(i))**:
   - Для каждого объекта вычисляется среднее расстояние до всех других объектов его кластера. Это внутрикластерное расстояние показывает, насколько близко объект находится к другим объектам своего кластера.
3. **Расчет межкластерного расстояния (b(i))**:
   - Для каждого объекта вычисляется минимальное среднее расстояние до объектов других кластеров. Это расстояние показывает, насколько далеко объект находится от ближайшего кластера.
4. **Вычисление индекса силуэта (s(i))**:
   - На основе значений a(i) и b(i) вычисляется индекс силуэта для каждого объекта по приведенной выше формуле.
5. **Визуализация индексов силуэта**:
   - Для интерпретации результатов силуэтного анализа строятся силуэтные графики. На силуэтных графиках отображаются значения индексов силуэта для каждого объекта, упорядоченные по возрастанию внутри каждого кластера.
   - Графики позволяют визуально оценить качество кластеров: чем шире и выше силуэты, тем лучше кластеризация.
6. **Анализ и интерпретация результатов**:

- Средний индекс силуэта по всем объектам используется для оценки общего качества кластеризации.
- Значения индексов силуэта анализируются для выявления объектов, которые могли быть неправильно кластеризированы.
- Выявляются кластеры с низкими значениями силуэтов, что может указывать на необходимость пересмотра параметров кластеризации или выбора другого метода.

Применение и результаты (Application and results)

1) Проведение кластеризации данных (Conducting data clustering)

1.2. Определение количества кластеров (Determining the number of clusters)

Для каждого метода кластеризации был построен график для визуального определения количества кластеров.

K-modes

Для определения оптимального количества кластеров применен метод локтя (Elbow Method).

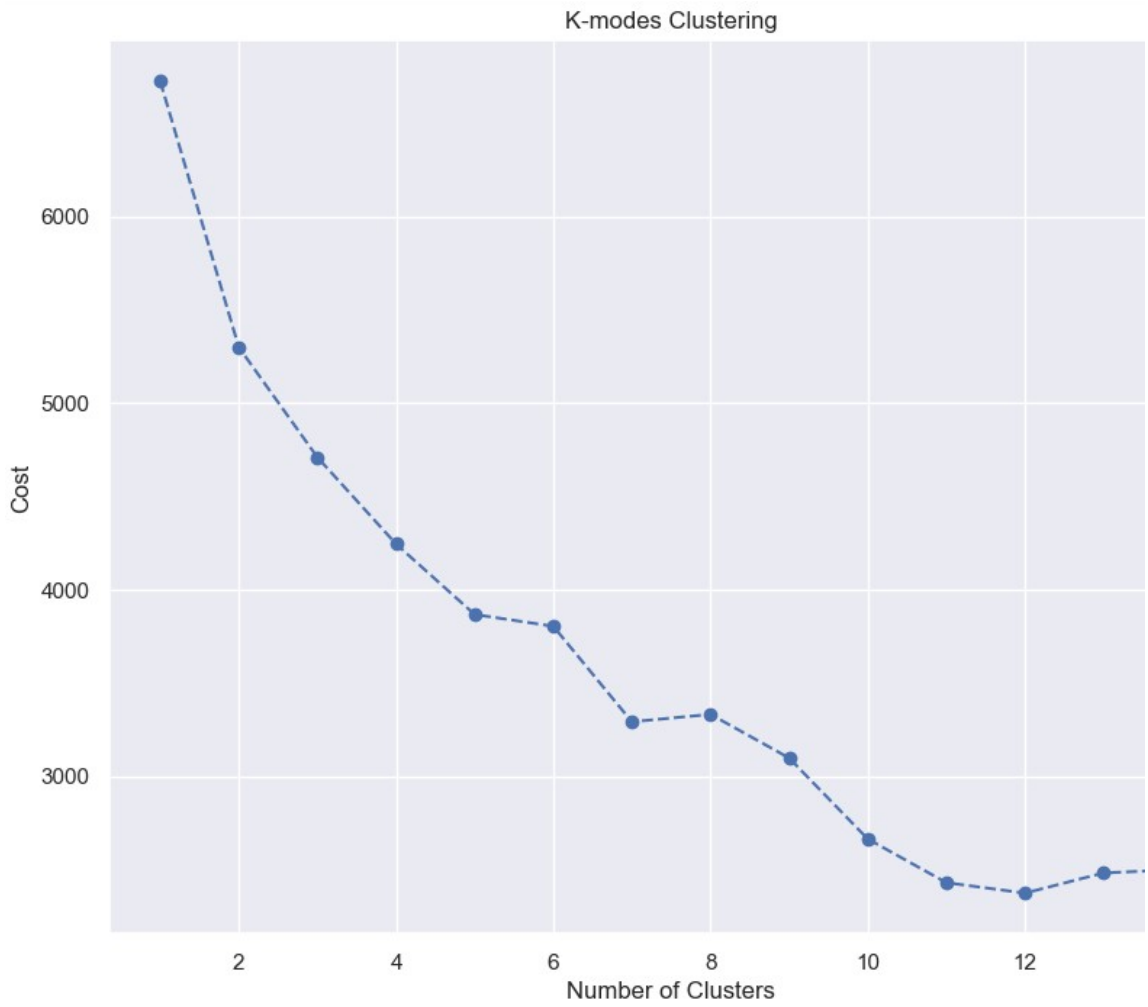

Рис. 3. График внутрикластерной суммы квадратов для разного количества кластеров

Pic. 3. Graph of intra-cluster sum of squares for different number of clusters

График не показал явно точку перелома. Поэтому по графику невозможно точно определить оптимальное количество кластеров.

Иерархическая кластеризация:

Оптимальное количество кластеров определялось по дендрограмме.

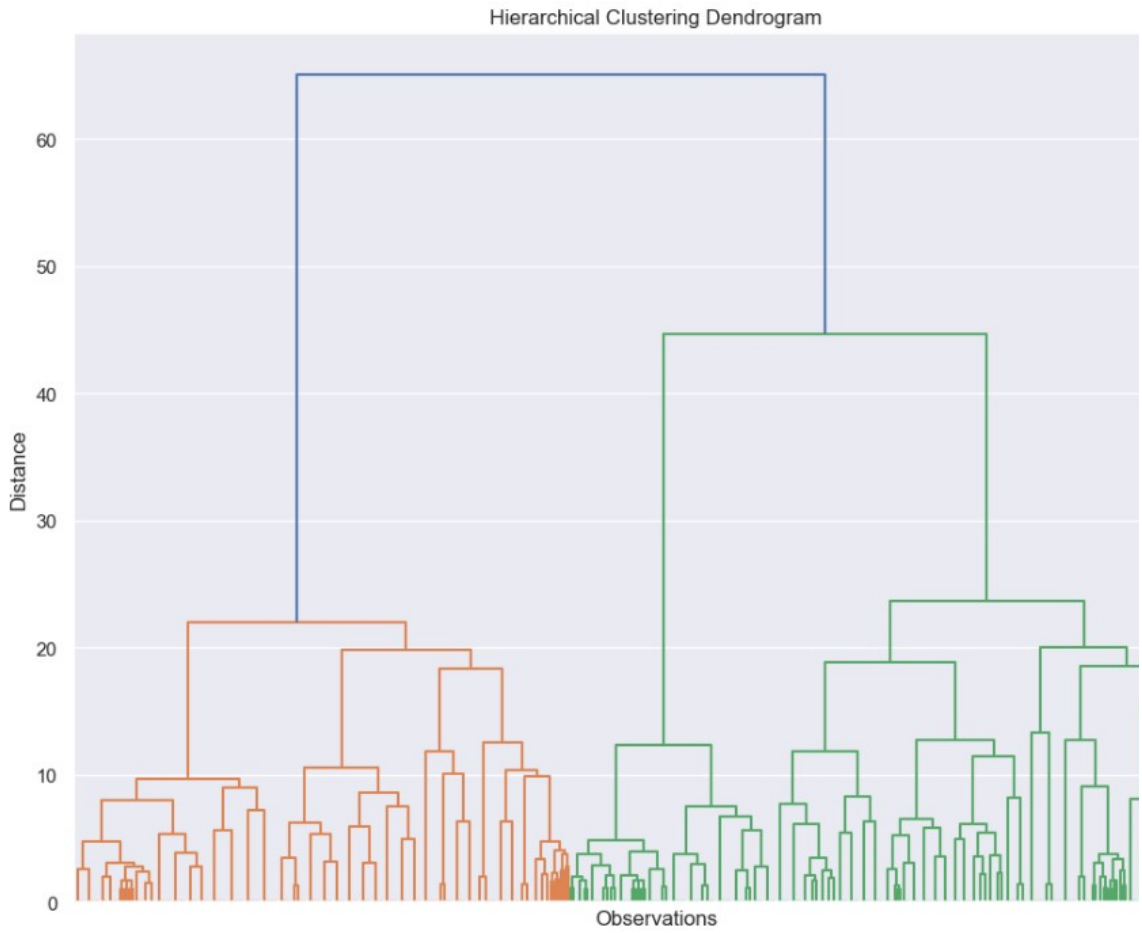

Рис. 4. График: Дендрограмма иерархической кластеризации

Pic. 4. Graph: Dendrogram of hierarchical clustering

Согласно цветовому коду на графике можно выделить всего 2 кластера, что предположительно мало для используемого объема и качества данных.

2) Расчет индексов силуэта (Calculation of silhouette indices)

2.1. Результаты расчета индексов силуэта для различных кластеров (Calculation results of silhouette indices for different clusters)

Визуальный анализ на предыдущем шаге четко не выявил точное количество необходимых кластеров. В связи с этим для каждого количества кластеров выполнен расчет индекса силуэта.

Для каждого метода кластеризации были рассчитаны индексы силуэта, что позволяет оценить качество полученных кластеров.

K-modes

Результаты расчета индекса силуэта для кластеров,

определенных методом K-modes

Для 2 кластеров индекс силуэта составляет: 0.17717677947063074

Для 3 кластеров индекс силуэта составляет: 0.10623737574862703

Для 4 кластеров индекс силуэта составляет: 0.09163491332494363

Для 5 кластеров индекс силуэта составляет: 0.15685845290616393

Для 6 кластеров индекс силуэта составляет: 0.14878518651548311

Для 7 кластеров индекс силуэта составляет: 0.239788928619284

Для 8 кластеров индекс силуэта составляет: 0.14057237325092672

Для 9 кластеров индекс силуэта составляет: 0.19965673529103695

Для 10 кластеров индекс силуэта составляет: 0.2682606002210953

Для 11 кластеров индекс силуэта составляет: 0.2333999111397604

Для 12 кластеров индекс силуэта составляет: 0.2708701682066994

Иерархическая кластеризация

Результаты расчета индекса силуэта для кластеров, определенных методом Иерархической кластеризации

Для 2 кластеров индекс силуэта составляет: 0.2843783849126773

Для 3 кластеров индекс силуэта составляет: 0.29980229889515286

Для 4 кластеров индекс силуэта составляет: 0.25585071394555176

Для 5 кластеров индекс силуэта составляет: 0.22614003862700519

Для 6 кластеров индекс силуэта составляет: 0.2169854330121357

Для 7 кластеров индекс силуэта составляет: 0.2360165463779702

Для 8 кластеров индекс силуэта составляет: 0.2613087487368942

Для 9 кластеров индекс силуэта составляет: 0.2834369890661361

Для 10 кластеров индекс силуэта составляет: 0.2954137516054588

Для 11 кластеров индекс силуэта составляет: 0.3132433742896329

Для 12 кластеров индекс силуэта составляет: 0.3234374363471928

2.2. Визуализация индексов силуэта (Visualization of silhouette indices)

Для каждого метода кластеризации были построены графики распределения индексов силуэтов в зависимости от количества кластеров, которые визуально представляют значения силуэтных коэффициентов для каждого кластера.

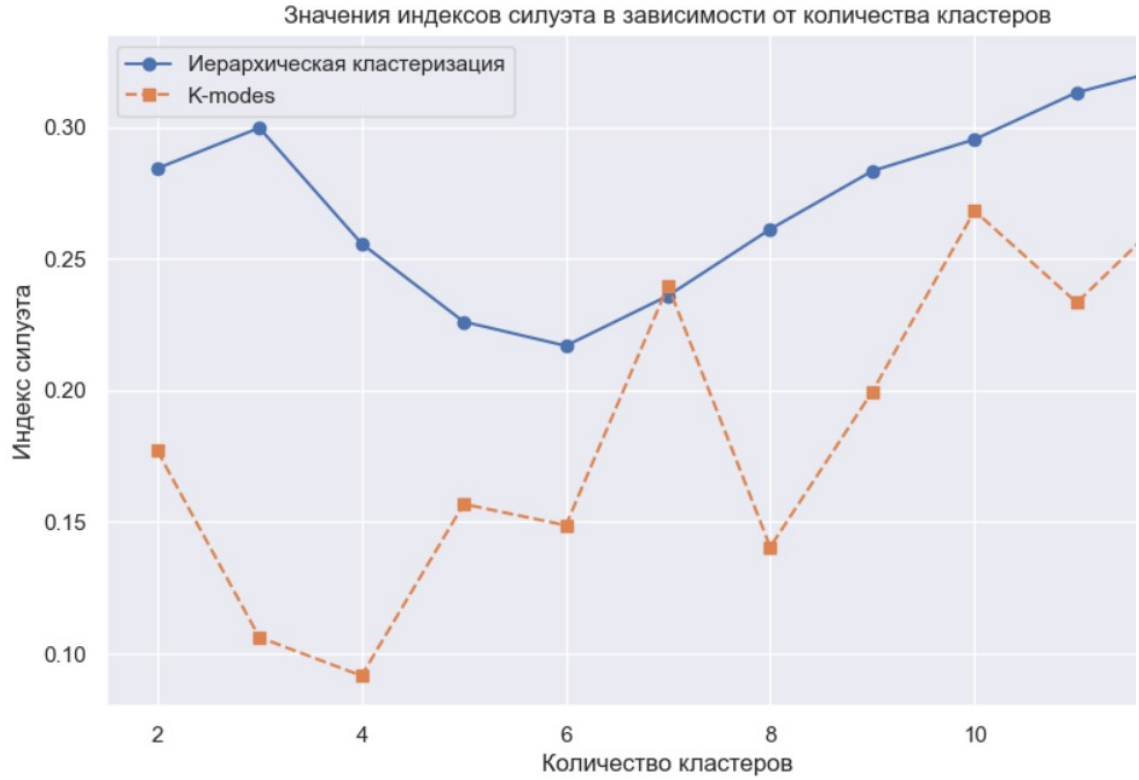

Рис. 5. График: Значения индексов силуэта в зависимости от количества кластеров

Pic. 5. Graph: Values of silhouette indices depending on the number of clusters

По полученному графику, который отображает значения индексов силуэта для иерархической кластеризации и метода K-modes в зависимости от количества кластеров, можно сделать следующие выводы:

K-modes:

- Индексы силуэта в целом ниже, что указывает на менее точное разбиение данных на кластеры этим методом.
- Начиная с 7 кластеров, индекс силуэта начинает улучшаться, и максимальные значения достигаются при 10 и 12 кластерах. Это может свидетельствовать о том, что данные начинают структурироваться более

явно при большем количестве кластеров.

Иерархическая кластеризация:

- Индексы силуэта для иерархической кластеризации выше, чем для K-modes, на всех этапах анализа, что говорит о лучшем качестве кластеризации в данном случае.
- Наблюдается увеличение индекса силуэта с ростом количества кластеров, особенно начиная с 10 кластеров. Максимум достигается при 12 кластерах, что может указывать на то, что именно это количество кластеров лучше всего описывает структуру данных.

3) Оценка качества кластеров (Quality assessment of clusters)

3.1. Сравнение результатов силуэтного анализа для различных методов кластеризации (Comparison of silhouette analysis results for different clustering methods)

Проведено сравнение результатов силуэтного анализа для всех методов кластеризации, что позволило определить наиболее эффективный метод для данного набора данных.

K-modes:

- Силуэтный индекс в этом случае ниже, чем в иерархической кластеризации, что указывает на то, что K-modes хуже сегментирует представленные данные.
- При этом заметен рост индекса при 7 и 10 кластерах, достигая максимума на 12 кластерах (0.270). Это свидетельствует о том, что сегментация на 12 кластеров наиболее оптимальна, хотя качество кластеризации всё ещё остается ниже по сравнению с иерархической кластеризацией.

Иерархическая кластеризация:

- Силуэтный индекс сначала показывает максимальное значение при 3 кластерах, а затем плавно снижается до 6 кластеров.

- Затем наблюдается небольшой рост индекса с увеличением числа кластеров до 12, где он достигает максимального значения (0.323).
- Это может говорить о том, что данные имеют естественные подгруппы, которые лучше всего выделяются при 11-12 кластерах.

3.2. Выявление оптимального метода кластеризации на основе силуэтного анализа (Identifying the optimal clustering method based on silhouette analysis)

На основе результатов силуэтного анализа было выявлено, что иерархическая кластеризация лучше описывает структуру данных, особенно при большем количестве кластеров (11-12 кластеров).

Метод K-modes показал хуже результаты по индексу силуэта, однако, похожая картина с улучшением на 12 кластерах предполагает, что это количество кластеров может быть оптимальным выбором для представленного набора данных.

Рекомендации:

- Несмотря на то, что метод K-modes рекомендуется для использования в задачах кластеризации категориальных данных, стоит рассмотреть комбинацию методов для подтверждения полученных выводов, а также обратить внимание на визуализацию кластеров для лучшего понимания качества разбиения
- Иерархическая кластеризация может быть использована для предварительного анализа структуры данных и определения количества кластеров.

Обсуждение (Discussion)

На основе проведенного исследования можно сделать несколько важных выводов о применении методов кластеризации K-modes и иерархической кластеризации, а также о роли силуэтного анализа в оценке их эффективности.

1) Анализ результатов и сравнительный анализ методов (Analysis of results and comparative analysis of methods)

Методы K-modes и иерархической кластеризации продемонстрировали различные результаты при применении к набору данных. Метод K-modes, предназначенный для работы с категориальными данными, показал себя как достаточно простой и эффективный инструмент. Однако его зависимость от предварительно заданного количества кластеров и чувствительность к выбору начальных центров кластеров стали основными ограничивающими факторами. С другой стороны, иерархическая кластеризация, не требующая заранее определения количества кластеров, позволила глубже понять структуру данных, что подтверждается лучшими значениями силуэтного индекса.

Тем не менее, следует отметить, что для большого количества данных иерархическая кластеризация оказалась более вычислительно сложной и чувствительной к шуму. Несмотря на это, она продемонстрировала более стабильные и высокие значения силуэтного индекса, что свидетельствует о лучшем качестве кластеризации по сравнению с методом K-modes.

2) Влияние количества кластеров (Effect of the number of clusters)

Исследование показало, что выбор оптимального количества кластеров является критическим фактором для успешной кластеризации. Силуэтный анализ подтвердил, что при меньшем количестве кластеров возможно недоразбиение данных, тогда как чрезмерное увеличение количества кластеров приводит к разбиению, которое не всегда отражает естественную структуру данных. Наиболее высокие значения силуэтного индекса были достигнуты при количестве кластеров, близком к 10-12, что и стало основным ориентиром для выбора числа кластеров в этом исследовании.

3) Практическая значимость (Practical Significance)

Силуэтный анализ оказался полезным инструментом для оценки качества кластеров, помогая выявить как удачные, так и проблемные

сегментации. В практическом применении, результаты данного исследования могут быть полезны в области маркетинга, где корректная сегментация клиентов позволяет улучшить целевую направленность маркетинговых кампаний.

4) Ограничения и перспективы дальнейших исследований (Limitations and prospects for further research)

Несмотря на полученные результаты, исследование имеет свои ограничения. Оно было проведено на одном наборе данных, что снижает возможность распространения выводов на другие типы данных. В будущем целесообразно рассмотреть применение других методов кластеризации, а также провести аналогичные исследования на других наборах данных.

Заключение (Conclusion)

В рамках данного исследования проведена оценка качества кластеров, сформированных с использованием двух методов кластеризации: K-modes и иерархической кластеризации, с применением силуэтного анализа. Результаты демонстрируют, что у каждого из методов есть свои преимущества и недостатки, и оптимальный выбор метода зависит от особенностей данных и цели исследования.

1) Основные выводы исследования (Key findings of the study)

Эффективность методов кластеризации:

- Метод K-modes продемонстрировал хорошие результаты при работе с категориальными данными, подтверждая свою полезность в задачах, где данные представлены в виде дискретных категорий. Основное преимущество этого метода заключается в его простоте и быстроте вычислений, что делает его применимым для больших наборов данных. Однако, его зависимость от предварительно заданного количества кластеров и чувствительность к выбору начальных центров накладывают определенные ограничения на его использование.

- Иерархическая кластеризация показала высокую эффективность в выделении естественных структур данных, особенно при работе с данными, которые могут быть сложно сегментированы заранее.[1] Этот метод не требует предварительного задания количества кластеров и позволяет получить визуальное представление структуры данных с помощью дендрограммы. Однако, вычислительная сложность метода и его чувствительность к шуму ограничивают его использование в условиях больших данных.

Роль силуэтного анализа:

- Силуэтный анализ зарекомендовал себя как надежный инструмент для оценки качества кластеризации. Он позволил объективно сравнить результаты различных методов и выбрать наиболее подходящий из них. Значения силуэтных коэффициентов предоставили количественные метрики, которые помогли выявить оптимальное количество кластеров и оценить качество разбиения данных.
- Одним из ключевых результатов исследования стало определение оптимального количества кластеров для каждого метода. Для K-modes оптимальное количество кластеров варьировалось между 10 и 12, что указывает на сложность структуры данных. Для иерархической кластеризации максимальные значения силуэтного индекса также были достигнуты при 12 кластерах, что подтверждает наличие естественных подгрупп в данных.

Практическая значимость результатов:

- Полученные результаты имеют значительное практическое значение в таких областях, как маркетинг и социальные науки. Правильная сегментация данных с использованием методов кластеризации и оценка их качества с помощью силуэтного анализа позволяют улучшить целевую направленность маркетинговых кампаний, точность моделей

[1] Introduction to Hierarchical Clustering with Python and Scikit-Learn. (n.d.). Retrieved from https://towardsdatascience.com/introduction-to-hierarchical-clustering-with-python-and-scikit-learn-5d7e5a56dee4

прогнозирования и понимание структуры данных в социальных исследованиях.

2) Рекомендации для практического применения методов кластеризации и силуэтного анализа (Recommendations for practical application of clustering and silhouette analysis methods)

На основании проведенного исследования можно дать следующие рекомендации для применения методов кластеризации и силуэтного анализа в реальных задачах:

Выбор метода кластеризации:

- Для работы с категориальными данными рекомендуется использовать метод K-modes, который является простым и эффективным для обработки больших объемов данных.
- Иерархическая кластеризация рекомендуется для предварительного анализа структуры данных и определения количества кластеров, особенно если данные имеют сложную, многомерную структуру.

Использование силуэтного анализа:

- Силуэтный анализ должен быть неотъемлемой частью процесса кластеризации, так как он предоставляет объективные метрики для оценки качества кластеров. Визуализация силуэтных коэффициентов поможет выявить проблемы в кластеризации и оптимизировать параметры модели.

Настройка параметров кластеризации:

- Настройку параметров, таких как количество кластеров, следует проводить с использованием силуэтного анализа и других метрик, таких как метод локтя или индекс Дэвиса-Булдина. Это позволит достичь наилучшего разбиения данных и улучшить интерпретируемость результатов.

3) Направления для дальнейших исследований (Directions for further research)

Для дальнейшего развития и углубления исследований в области кластеризации и силуэтного анализа предлагаются следующие направления:

Расширение набора данных: Проведение аналогичных исследований на более разнообразных и крупных наборах данных для подтверждения обобщаемости полученных выводов.

Исследование других методов кластеризации: Анализ дополнительных методов, таких как спектральная кластеризация или метод самоорганизующихся карт (SOM), которые могут предложить альтернативные подходы к сегментации данных.

Комбинирование метрик оценки качества: Использование различных метрик для комплексного анализа качества кластеров, что позволит более точно оценивать результаты кластеризации и улучшать качество принимаемых решений.

Автоматизация процесса кластеризации: Разработка автоматизированных методов настройки параметров кластеризации, которые позволят значительно упростить процесс и повысить точность разбиения данных на кластеры.

Таким образом, проведенное исследование вносит значимый вклад в развитие методов кластеризации и силуэтного анализа, предлагая практические рекомендации для их применения и обозначая направления для дальнейших исследований.

Источники

1. Тан, П., Штайнбах, М., Кумар, В. Введение в методы интеллектуального анализа данных. Москва: Вильямс, 2014.
2. Гудфеллоу, И., Бенджио, Й., Курвилль, А. Глубокое обучение. Москва: ДМК Пресс, 2018.
3. Джеймс, Г., Виттен, Д., Хасти, Т., Тибширани, Р. Введение в статистическое обучение с примерами на языке R. Москва: ДМК Пресс, 2017.
4. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics, 1987. Т. 20, № 1. С. 53-65.

5. Jain, A. K., Murty, M. N., Flynn, P. J. Data clustering: a review // ACM Computing Surveys, 1999. Т. 31, № 3. С. 264-323. DOI: если доступен.
6. Xu, R., Wunsch, D. Survey of clustering algorithms // IEEE Transactions on Neural Networks, 2005. Т. 16, № 3. С. 645-678.
7. Estivill-Castro, V. Why so many clustering algorithms: A position paper // ACM SIGKDD Explorations Newsletter, 2002. Т. 4, № 1. С. 65-75. DOI: 10.1145/568574.568575.
8. Kaufman, L., Rousseeuw, P. J. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics, 1990.
9. Scikit-learn: Machine Learning in Python. [Электронный ресурс]. URL: https://scikit-learn.org/stable/ (дата обращения: 15.08.2024).
10. K-means and K-modes: A Comparison. [Электронный ресурс]. URL: https://www.datascience.com/blog/k-means-and-k-modes-clustering (дата обращения: 15.08.2024).
11. Introduction to Hierarchical Clustering with Python and Scikit-Learn. [Электронный ресурс]. URL: https://towardsdatascience.com/introduction-to-hierarchical-clustering-with-python-and-scikit-learn-5d7e5a56dee4 (дата обращения: 15.08.2024).
12. Towards Data Science. [Электронный ресурс]. URL: https://towardsdatascience.com/ (дата обращения: 15.08.2024).
13. Stack Overflow. [Электронный ресурс]. URL: https://stackoverflow.com/ (дата обращения: 15.08.2024).
14. Coursera: Machine Learning. [Электронный ресурс]. URL: https://www.coursera.org/learn/machine-learning (дата обращения: 15.08.2024).
15. YouTube: Data School. [Электронный ресурс]. URL: https://www.youtube.com/user/dataschool (дата обращения: 15.08.2024).