DOI: 10.58168/MIST2026_309-314
УДК 004

КАК ПРАВИЛЬНО СЕГМЕНТИРОВАТЬ ПОЛЬЗОВАТЕЛЕЙ: СРАВНИВАЕМ АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ ДЛЯ БИЗНЕС-ЗАДАЧ

Д.С. Кукуева[1], А.М. Тюнина[1], В.В. Кондусова[1]

[1]ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

Аннотация. В статье проводится сравнение трех алгоритмов кластеризации - K-Means, DBSCAN и иерархической кластеризации - для решения задачи сегментации пользователей цифровых сервисов. На основе экспериментов с синтетическими и реальными данными оценивается их эффективность, устойчивость к шуму, скорость работы и интерпретируемость результатов. Показано, что выбор алгоритма зависит от структуры данных и бизнес-целей, а ключевым фактором успеха является качественная предобработка данных.

Ключевые слова: сегментация пользователей, кластерный анализ, K-Means, DBSCAN, иерархическая кластеризация, машинное обучение без учителя, поведенческий анализ, анализ данных, сравнительный анализ алгоритмов

HOW TO SEGMENT USERS CORRECTLY: COMPARING CLUSTERING ALGORITHMS FOR BUSINESS TASKS

D.S. Kukueva[1], A.M. Tyunina[1], V.V. Kondusova[1]

[1]Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. The article compares three clustering algorithms - K-Means, DBSCAN and hierarchical clustering - to solve the problem of segmentation of users of digital services. Based on experiments with synthetic and real data, their effectiveness, noise tolerance, speed of operation, and interpretability of the results are evaluated. It is shown that the choice of an algorithm depends on the data structure and business goals, and high-quality data preprocessing is a key success factor.

Keywords:user segmentation, cluster analysis, K-Means, DBSCAN, hierarchical clustering, unsupervised machine learning, behavioral analysis, data analysis, comparative analysis of algorithms

© Кукуева Д. С., Тюнина А. М., Кондусова В. В., 2026

Эра цифровой экономики перенесла центр конкуренции из плоскости функциональности продукта в плоскость качества пользовательского опыта. Современные потребители ожидают от сервисов не просто исполнения базовых функций, но и глубокой персонализации, которая проявляется в релевантном контенте, своевременных предложениях и адаптивном интерфейсе. Основой для построения таких интеллектуальных систем является сегментация - разделение неоднородной массы пользователей на однородные группы (сегменты) со схожими характеристиками, потребностями и моделями поведения.

Традиционные методы сегментации, основанные на демографических или географических признаках, уступают в эффективности поведенческому анализу, который оперирует цифровыми следами пользователя: историей посещений, паттернами взаимодействия с интерфейсом, предпочтениями в контенте. Обработка этих больших, многомерных и зашумленных массивов данных требует применения формальных математических методов, среди которых кластерный анализ занимает центральное место. Кластеризация, как задача машинного обучения без учителя, позволяет обнаружить скрытые структуры и закономерности в данных без привлечения заранее размеченных примеров.

Однако разнообразие алгоритмов кластеризации и их принципиально разные теоретические основания ставят перед аналитиком сложный выбор. Наивное применение наиболее популярного метода K-Means к данным сложной структуры может привести к методически неверным и практически бесполезным результатам. Таким образом, возникает необходимость в системном сопоставлении, которое бы учитывало не только абстрактное качество кластеризации, но и операциональные характеристики алгоритмов в условиях, максимально приближенных к реальным бизнес-задачам.

Целью данной работы является проведение сравнительного анализа трех широко распространенных алгоритмов кластеризации - K-Means, DBSCAN и иерархической кластеризации - для решения задачи сегментации пользователей цифровых сервисов. Исследование ставит перед собой задачи: 1) оценить способность алгоритмов корректно восстанавливать кластеры разной формы на синтетических данных; 2) протестировать их эффективность на реальных поведенческих данных; 3) сравнить по ключевым практико-ориентированным критериям; 4) сформулировать обоснованные рекомендации по выбору алгоритма в зависимости от специфики данных и бизнес-целей.

K-Means - итеративный центроидный алгоритм, целью которого является минимизация суммарного квадрата расстояний от точек кластера до его центра

(инерции). Пользователь задает параметр K - количество кластеров. Алгоритм начинает со случайной инициализации центров, затем на каждом шаге присваивает точки ближайшему центру и пересчитывает положение центров как среднее арифметическое точек кластера. Процесс продолжается до стабилизации. Его сильные стороны - простота, скорость и масштабируемость. Слабые - чувствительность к начальной инициализации, необходимость заранее задавать K, предположение о сферичности и одинаковом размере кластеров, высокая уязвимость к выбросам. [1]

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) - алгоритм, основанный на понятии плотности. Он не требует задания числа кластеров, вместо этого использует два параметра: ε (эпсилон) - радиус окрестности, и minPts - минимальное количество точек в этой окрестности. Точки делятся на три типа: ядерные (имеющие не менее minPts соседей в ε-окрестности), достижимые (находящиеся в окрестности ядерной) и шумовые. Кластер формируется из плотно связанных ядерных и достижимых точек. Ключевые преимущества: способность находить кластеры произвольной формы, устойчивость к выбросам, автоматическое определение числа кластеров. Недостатки: сложность выбора параметров для данных с разной плотностью, снижение эффективности в высокоразмерных пространствах. [2]

Иерархическая кластеризация (агломеративный подход) - алгоритм, строящий древовидную структуру (дендрограмму), отражающую вложенность кластеров. Начинается с того, что каждая точка считается отдельным кластером. На каждом шаге два ближайших кластера объединяются. Процесс продолжается, пока все точки не окажутся в одном кластере. Расстояние между кластерами может вычисляться разными способами (одиночная, полная, средняя связь). Главное достоинство - наглядность дендрограммы, позволяющая выбрать уровень детализации и увидеть иерархию данных. Главный недостаток - высокая вычислительная сложность, делающая метод неприменимым к очень большим наборам данных.

Для всесторонней оценки были использованы два типа данных.

1. Синтетические данные. Сгенерированы с использованием библиотеки `scikit-learn` для создания контролируемых сценариев:

2. Реальные данные. Анонимизированная выборка из 15 000 активных пользователей медиасервиса за квартал. Признаки были предобработаны (нормализация, обработка пропусков) и включали:

Эксперименты на синтетических наборах наглядно выявили сильные и слабые стороны каждого подхода.

На наборе A (сферические кластеры) все три алгоритма показали хорошие результаты. K-Means продемонстрировал наивысший ARI (0.99), идеально восстановив кластеры. DBSCAN, при правильно подобранном ε, также достиг ARI 0.98, корректно отнеся шумовые точки в отдельную категорию. Иерархическая кластеризация (со средней связью) показала ARI 0.97. На этом простом примере K-Means был самым быстрым.

На наборе B (полумесяцы) результаты радикально разошлись. K-Means полностью провалился (ARI ~0.1), безуспешно пытаясь разделить данные на сферические группы. DBSCAN блестяще справился с задачей (ARI 0.99), четко выделив оба невыпуклых кластера и весь шум. Иерархическая кластеризация показала промежуточный, но неудовлетворительный результат (ARI ~0.6), так как на ранних шагах начала объединять точки из разных, но близких в пространстве полумесяцев. [3]

На наборе C (концентрические кольца) ситуация повторилась. K-Means и иерархический метод оказались неспособны разделить кольца, так как основываются на евклидовой близости, а точки соседних колец близки друг к другу. DBSCAN вновь показал превосходный результат, корректно идентифицировав три отдельных кольца как плотностные кластеры.

Применение алгоритмов к реальным данным потребовало дополнительных шагов по настройке и интерпретации.

K-Means был запущен с K=5, определенным по методу локтя. Время выполнения составило менее 2 секунд. Силуэтный коэффициент равнялся 0.52, что указывает на среднее качество разделения. Анализ центроидов позволил выделить сегменты: «Эпизодические пользователи», «Регулярные читатели новостей», «Вечерние развлекатели». Однако алгоритм «размазал» небольшую группу сверхактивных пользователей по нескольким кластерам, исказив их профили. [4]

DBSCAN (с параметрами ε=0.3, minPts=15) выделил 4 плотных кластера и отнес около 8% точек к шуму. Силуэтный коэффициент для основных кластеров был выше - 0.61. Выделенные группы оказались более однородными и чистыми: например, четко обособился кластер «Спортивные фанаты на десктопе» с высокой концентрацией на спортивном контенте и активностью только с компьютеров. Группа, отнесенная к шуму, при ручном анализе оказалась смесью технических ботов и уникальных пользователей с нетипичным поведением, что является

ценной находкой для службы безопасности. Основным недостатком стала длительная ручная настройка параметров.

Иерархическая кластеризация (метод Уорда) оказалась самой медленной (около 90 секунд). По дендрограмме был выбран уровень отсечения, давший 6 кластеров. Силуэтный коэффициент был низким (0.44), что свидетельствует о плохой разделимости полученных групп. Однако визуализация в виде дендрограммы позволила легко идентифицировать два крупных макро-сегмента («Пассивная аудитория» и «Активная аудитория») и увидеть их внутреннюю структуру, что ценно для стратегического анализа.

Проведенный сравнительный анализ убедительно доказывает, что не существует «лучшего» алгоритма кластеризации в абсолютном смысле. Его выбор является важным проектировочным решением, которое должно основываться на глубоком понимании природы данных и четко сформулированной бизнес-цели.

Таким образом, процесс сегментации следует рассматривать как комплексную аналитическую цепочку, где выбор алгоритма кластеризации является одним из завершающих, а не стартовых шагов. Инвестиции времени в исследование данных, их очистку и конструирование признаков обеспечивают большую отдачу, чем поиск мифического «самого точного» алгоритма. Комбинация разных методов (например, использование DBSCAN для очистки данных от шума с последующей кластеризацией очищенного набора с помощью K-Means) часто дает наилучший практический результат, позволяя строить robust и интерпретируемые модели пользовательских сегментов.

Список литературы

1. Волкова, Е. Е. Особенности поведения российских потребителей в возрасте 18-23 лет на рынке цифрового музыкального контента / Е. Е. Волкова // Hypothesis. – 2018. – № 2(3). – С. 33-40. – EDN LIWTSS.

2. Прокопенко, Н. Ю. Применение Loginom для оптимизации процессов управления товарными запасами предприятий малого и среднего бизнеса / Н. Ю. Прокопенко, Д. В. Тришин // Межвузовский сборник статей лауреатов конкурсов : Сборник статей / Нижегородский государственный архитектурно-строительный университет. Том ВЫПУСК 21. – Нижний Новгород : Нижегородский государственный архитектурно-строительный университет, 2021. – С. 216-222. – EDN TLNXXH.

3. Витяев, Е. Е. Семантический вероятностный вывод предсказаний / Е. Е. Витяев // Известия Иркутского государственного университета. Серия: Математика. – 2017. – Т. 21. – С. 33-50. – DOI 10.26516/1997-7670.2017.21.33. – EDN ZFOTBF.

4. Сумин, В. И. Особенности выбора членов экспертной группы для анализа функционирования сложной организационной системы силовых структур / В. И. Сумин, А. С. Дубровин, И. С. Кущева // Моделирование систем и процессов. – 2024. – Т. 17, № 4. – С. 77-83. – DOI 10.12737/2219-0767-2024-17-4-77-83. – EDN ISTAML.

References

1. Volkova, E. E. Recognition of the ruler of Russia at the age of 18-23 in the music content market / E. E. Volkova // Hypothesis. – 2018. – № 2(3). – Pp. 33-40. – EDN LIVTS.

2. Prokopenko, N. Y. The use of login for optimizing inventory management processes of small and medium-sized businesses / N. Y. Prokopenko, D. V. Trishin // Interuniversity collection of articles by laureates of competitions : Collection of articles / Nizhny Novgorod State University of Architecture and Civil Engineering. VOLUME ISSUE 21. Nizhny Novgorod : Nizhny Novgorod State University of Architecture and Civil Engineering, 2021. pp. 216-222. EDN TLNXXXX.

3. Vityaev, E. E. Semantic probabilistic prediction inference / E. E. Vityaev // Izvestiya Irkutsk State University. Series: Mathematics. – 2017. – Vol. 21. – pp. 33-50. – DOI 10.26516/1997-7670.2017.21.33. – ZFOTBF PUBLISHING HOUSE.

4. Sumin, V. I. Features of the selection of members of the expert group for the analysis of the functioning of a complex organizational system of law enforcement agencies / V. I. Sumin, A. S. Dubrovin, I. S. Kushcheva // Modeling of systems and processes. – 2024. – Vol. 17, No. 4. – pp. 77-83. – In doi 10.12737/2219-0767-2024-17-4-77-83. – EDN ISTAML.