УДК 004.8:004.6

Бессонов Николай Антонович
студент
Санкт-Петербургский государственный университет аэрокосмического приборостроения

ПРИМЕНЕНИЕ АЛГОРИТМА КЛАСТЕРИЗАЦИИ K-MEANS ДЛЯ СЕГМЕНТАЦИИ ПОЛЬЗОВАТЕЛЕЙ В СИСТЕМАХ АНАЛИЗА ИНФОРМАЦИОННЫХ РЕСУРСОВ ИНТЕРНЕТ-ПЛАТФОРМ

Abstract. The paper considers the application of the k-means clustering algorithm for user segmentation in internet platform information resource analysis systems. The mathematical model of minimizing intra-cluster variance is presented. The advantages and limitations of the method in processing large-scale behavioral datasets are analyzed. Practical examples of using k-means in digital marketing, recommendation systems and user analytics are discussed.

Keywords: k-means, clustering, user segmentation, data analysis, internet platforms, information resources, unsupervised learning.

Аннотация

Рассматривается задача сегментации пользователей интернет-платформ на основе алгоритма кластеризации k-means. Представлена математическая модель минимизации внутрикластерной дисперсии, описаны этапы работы алгоритма и особенности его применения при анализе больших массивов пользовательских данных. Приведены примеры практического использования метода в системах цифрового маркетинга, рекомендательных сервисах и аналитике поведения пользователей.

Ключевые слова: кластеризация, k-means, сегментация пользователей, анализ данных, интернет-платформы, информационные ресурсы, машинное обучение.

Введение

В условиях цифровой трансформации общества интернет-платформы становятся важнейшими центрами накопления и обработки информационных ресурсов. Социальные сети, маркетплейсы, образовательные порталы и медиасервисы ежедневно генерируют значительные объемы пользовательских данных. Эффективное управление такими ресурсами невозможно без применения интеллектуальных методов анализа данных. Современные организации сталкиваются с необходимостью оперативного выявления

закономерностей в поведении пользователей, прогнозирования их потребностей и адаптации сервисов под индивидуальные предпочтения. В этом контексте аналитические системы, использующие методы машинного обучения, становятся ключевым инструментом повышения эффективности бизнес-процессов и улучшения качества предоставляемых услуг [1].

Одной из ключевых задач является сегментация пользователей — процесс разбиения аудитории на однородные группы по поведенческим, социально-демографическим и транзакционным признакам. Сегментация позволяет повысить точность персонализации контента, оптимизировать маркетинговые стратегии, улучшить пользовательский опыт и снизить издержки на привлечение и удержание аудитории. Кроме того, она обеспечивает основу для разработки рекомендательных систем, программ лояльности и целевых рекламных кампаний [2].

Для решения данной задачи широко применяются методы машинного обучения, в частности алгоритмы неконтролируемого обучения. Одним из наиболее распространенных методов является алгоритм k-means, предложенный J. MacQueen в 1967 году [3]. Его популярность обусловлена вычислительной эффективностью и относительной простотой реализации. K-means позволяет автоматически выделять группы пользователей с похожими характеристиками, что особенно важно при работе с большими объемами данных, где ручной анализ становится невозможным. Алгоритм может эффективно применяться как в коммерческих, так и в научных исследованиях, обеспечивая аналитическую поддержку при принятии управленческих решений и формировании стратегии развития интернет-платформ.

Теоретические основы алгоритма k-means

Алгоритм k-means относится к методам разбиения множества объектов на заранее заданное число кластеров k. Основной целью алгоритма является минимизация суммарного квадратичного отклонения объектов от центров соответствующих кластеров.

Математическая модель алгоритма описывается функцией:

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2,$$

где:

S_i - множество объектов i-го кластера;

μ_i - центр (среднее значение) i-го кластера;

$\|x - \mu_i\|^2$- квадрат евклидова расстояния между объектом и центром кластера [4].

Минимизация функции J обеспечивает формирование компактных и максимально однородных кластеров. Алгоритм начинается с задания числа кластеров k и инициализации центров кластеров. Далее каждый объект назначается к ближайшему центру по выбранной метрике расстояния, после чего выполняется пересчет центров кластеров. Эти шаги повторяются до достижения сходимости, то есть пока изменения центров кластеров становятся минимальными. Алгоритм характеризуется линейной вычислительной сложностью по числу объектов, что делает его пригодным для анализа больших данных [5].

Применение k-means для сегментации пользователей

В системах анализа информационных ресурсов интернет-платформ алгоритм k-means применяется для группировки пользователей по сходству их цифрового поведения. В качестве признакового пространства обычно используются количественные характеристики, отражающие активность пользователя: частота посещений платформы, продолжительность сессий, временные интервалы активности, количество и объем транзакций, глубина взаимодействия с контентом, предпочтительные категории материалов и иные параметры, формирующие поведенческий профиль. На основе этих данных формируются кластеры, объединяющие пользователей с близкими характеристиками, что позволяет перейти от индивидуального анализа к анализу сегментов аудитории [6].

В цифровом маркетинге метод используется для выделения целевых групп потребителей, что дает возможность формировать персонализированные рекламные кампании и адаптировать коммуникационные стратегии под особенности каждого сегмента. Кластеризация способствует более точному позиционированию продукта и рациональному распределению маркетингового бюджета. В рекомендательных системах предварительная сегментация пользователей повышает релевантность рекомендаций за счет учета коллективных интересов схожих групп. Пользователи, входящие в один кластер, с высокой вероятностью демонстрируют близкие предпочтения, что позволяет строить модели рекомендаций на основе агрегированных характеристик сегмента [5].

В научных исследованиях отмечается, что сочетание RFM-анализа (Recency, Frequency, Monetary) с алгоритмом k-means обеспечивает эффективное выявление наиболее ценных сегментов клиентов. Такой подход позволяет учитывать не только поведенческие, но и экономические показатели, характеризующие вклад пользователя в доход платформы. В отечественных работах также подтверждается эффективность метода при анализе больших массивов данных в CRM-системах, где кластеризация используется для

автоматизации взаимодействия с клиентами, прогнозирования оттока и повышения уровня удержания аудитории. Таким образом, применение алгоритма k-means позволяет выявлять группы лояльных пользователей, определять сегменты с высоким потенциалом монетизации, анализировать устойчивые поведенческие паттерны и в целом повышать эффективность управления информационными ресурсами интернет-платформ [7].

Преимущества и ограничения метода

Алгоритм k-means обладает рядом существенных преимуществ, обеспечивающих его широкое распространение в аналитических системах. Прежде всего, он характеризуется высокой скоростью работы и сравнительно невысокой вычислительной сложностью, что особенно важно при обработке больших объемов пользовательских данных. Простота математической модели и прозрачность логики функционирования делают алгоритм удобным для практической реализации и интерпретации результатов. Метод хорошо масштабируется и может применяться в распределенных вычислительных средах. Центры кластеров формируют обобщенные профили типичных представителей сегментов, что облегчает аналитическую интерпретацию полученных групп [2].

Вместе с тем алгоритм имеет ряд ограничений, которые необходимо учитывать при его использовании. Одним из ключевых является необходимость предварительного задания числа кластеров, что требует проведения дополнительного анализа или применения специальных критериев выбора оптимального значения k. Результаты кластеризации могут зависеть от выбора начальных центров, поскольку алгоритм сходится к локальному минимуму целевой функции. Метод также чувствителен к выбросам и аномальным значениям, способным существенно смещать положение центроидов. Кроме того, алгоритм предполагает приблизительно сферическую форму кластеров и использование евклидовой метрики расстояния, что не всегда адекватно отражает сложную структуру реальных данных. Для повышения устойчивости результатов на практике применяются усовершенствованные методы инициализации, такие как k-means++, а также выполняются множественные запуски алгоритма с выбором наилучшего решения по значению целевой функции [4].

Заключение

Алгоритм k-means является эффективным инструментом сегментации пользователей в системах анализа информационных ресурсов интернет-платформ. Его применение позволяет структурировать большие массивы пользовательских данных, выявлять

закономерности поведения аудитории и формировать аналитическую основу для принятия управленческих решений. Метод обеспечивает баланс между вычислительной эффективностью и качеством получаемых результатов, что делает его одним из базовых инструментов анализа данных в задачах цифровой экономики.

Несмотря на существующие ограничения, k-means продолжает активно использоваться в практических и научных исследованиях благодаря своей универсальности и простоте адаптации к различным предметным областям. Перспективным направлением развития является интеграция алгоритма с более сложными моделями машинного обучения, применение гибридных методов кластеризации, а также использование автоматизированных процедур определения оптимального числа кластеров. Это позволит повысить точность сегментации и обеспечить более гибкую адаптацию аналитических систем к динамично изменяющейся среде интернет-платформ.

Библиографический список

[1] Д. А. Клинов и К. А. Григорян, «Разработка методики сегментации пользователей с помощью алгоритмов кластеризации и расширенной аналитики», *Электронные библиотеки*, т. 25, вып. 2, сс. 137–147, апр. 2022, doi: 10.26907/1562-5419-2022-25-2-137-147.

[2] А. Платонова и М. Рыжкова, «Применение машинного обучения для сегментации пользователей в маркетинговых исследованиях», *Радиотехнические и телекоммуникационные системы*, вып. 3, сс. 47–53, 2025.

[3] J. MacQueen, «Some methods for classification and analysis of multivariate observations», в *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, т. 5.1, University of California Press, 1967, сс. 281–298. Просмотрено: 1 март 2026 г. [Онлайн]. Доступно на: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

[4] A. K. Jain, «Data clustering: 50 years beyond K-means», *Pattern Recognition Letters*, т. 31, вып. 8, сс. 651–666, июн. 2010, doi: 10.1016/j.patrec.2009.09.011.

[5] K. Tabianan, S. Velu, и V. Ravi, «K-means clustering approach for intelligent customer segmentation using customer purchase behavior data», *Sustainability*, т. 14, вып. 12, июн. 2022, doi: 10.3390/su14127243.

[6] М. Сычев, А. Астрахов, Д. Правиков, и О. Тягунков, «Применение методов кластеризации для анализа неиндексируемых интернет-ресурсов», *Инженерный журнал: наука и инновации*, вып. 2 (14), с. 8, 2013.

[7] F. M. Talaat, A. Aljadani, B. Alharthi, M. A. Farsi, M. Badawy, и M. Elhosseini, «A mathematical model for customer segmentation leveraging deep learning, explainable AI, and RFM analysis in targeted marketing», *Mathematics*, т. 11, вып. 18, сен. 2023, doi: 10.3390/math11183930.