УДК 339:004

ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАСТЕРИЗАЦИИ КЛИЕНТОВ В ЭЛЕКТРОННОЙ КОММЕРЦИИ

М. Д. Мацуганова, Д. В. Черненко
Витебский государственный технологический университет, Беларусь

Рассмотрены основные методы кластеризации – группировки клиентов по схожим признакам, используемой в алгоритмах машинного обучения при анализе поведения клиентов в электронной коммерции.

Электронная коммерция стала неотъемлемой частью мировой розничной торговли. Как и многие другие отрасли, купля-продажа товаров претерпела существенные изменения с появлением Интернета, и благодаря продолжающейся цифровизации современной жизни потребители по всему миру теперь могут воспользоваться преимуществами онлайн-транзакций. По мере стремительного роста доступа к Интернету и его внедрения число людей, совершающих покупки онлайн, постоянно растет. По оценкам, в 2025 г. объем розничных продаж в сфере электронной коммерции по всему миру превысит 4,3 трлн долл. США и ожидается, что дальнейший рост составит до 10 трлн к 2033 г. [1]. С развитием онлайн-торговли объемы собираемых данных стремительно увеличиваются.

Такое многообразие и объем данных открывают новые возможности для бизнеса, но одновременно требуют более сложных и интеллектуальных методов обработки. Именно здесь на первый план выходит машинное обучение – область знаний, изучающая методы построения алгоритмов, способных автоматически улучшать свое поведение на основе накопленного опыта и данных [2].

Благодаря алгоритмам машинного обучения компании могут: формировать персонализированные рекомендации в реальном времени, оптимизировать ценообразование и автоматизировать сегментацию аудитории. Таким образом, переход от количественного роста продаж к качественной работе с данными становится ключевым фактором конкурентоспособности в цифровой экономике.

Рассмотрим сегментацию клиентов подробнее. Сегментация и кластеризация клиентов – это два ключевых подхода к анализу аудитории, позволяющие компаниям глубже понимать поведение потребителей и выстраивать более точные стратегии взаимодействия. Сегментация представляет собой процесс разделения клиентов на группы по заранее заданным признакам, таким как демография, география, поведение.

Кластеризация, в отличие от сегментации, основана на алгоритмах машинного обучения и не требует заранее определенных критериев [3]. К наиболее популярным алгоритмам машинного обучения можно отнести k-means и DBSCAN.

Работа алгоритма k-means начинается со случайного выбора k центров кластеров, где k – это заранее заданное количество групп. Далее каждый объект из набора данных присваивается тому кластеру, центр которого находится ближе всего согласно выбранной метрике расстояния. После назначения всех объектов центры кластеров пересчитываются как среднее значение координат всех точек, входящих в соответствующий кластер. Шаги (перераспределение объектов и обновление центров) повторяются итеративно до тех пор, пока положения центров не стабилизируются или не будет достигнуто заданное число итераций [4].

Для эффективного внедрения алгоритма необходим достоверный и разнообразный массив данных, обеспечивающий высокое качество работы алгоритмов машинного обучения, но до начала работы алгоритма k-means должен быть осуществлен сбор данных и предварительная обработка, которая включает в себя очистку данных, интеграцию и преобразование.

Наиболее распространенным источником информации для анализа являются транзакционные данные, которые обычно включают следующие ключевые параметры: *customer id, orders, avg_spend и discount_usage.* Данные параметры были выбраны за их высокую информативность и практическую значимость для анализа поведения покупателей. Customer_id – это уникальный идентификатор клиента, который не участвует напрямую, но необходим для отслеживания пользователей. Orders (количество заказов) отражает активность клиента. Чем больше заказов, тем выше вовлеченность и лояльность. Avg_spend (средний чек) показывает финансовую ценность клиента. Discount usage (частота использования скидок) характеризует чувствительность клиента к акциям и промопредложениям. Это важно для определения того, как клиенты реагируют на скидки. Пример реализации алгоритма k-means показан на рис. 1.

```
# Масштабирование признаков
features = ['orders', 'avg_spend', 'discount_usage']
scaler = Standardscaler()
scaled_data = scaler.fit_transform(df[features])
# Кластеризация K-means
kmeans = KMeans (n_clusters=3, random_state=42)
df['cluster'] kmeans.fit_predict(scaled_data)
```

Рис. 1. Пример реализации алгоритма k-means

После применения алгоритма k-means к данным о клиентах сформировались несколько групп (кластеров), каждая из которых отражает определенный тип поведения. Для визуализации работы алгоритма был создан график (рис. 2).

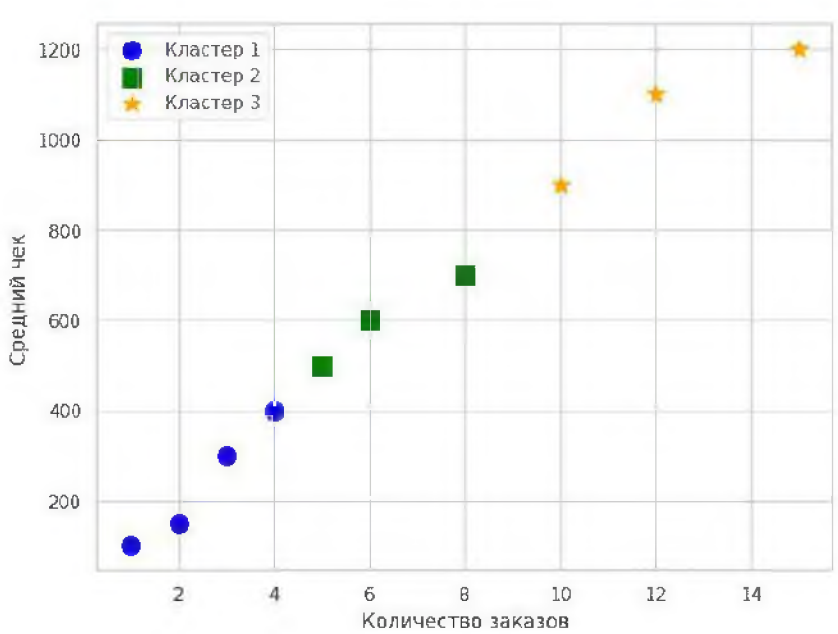

Рис. 2. Кластеризация с использованием алгоритма k-means

Таким образом, можно сделать вывод, что алгоритм автоматически разделил клиентов на кластеры, каждый из которых характеризуется определенным стилем покупок. Кластер 3 включает наиболее ценных клиентов. Они доверяют бренду, покупают регулярно. Их стоит удерживать через программы лояльности, эксклюзивные предложения и персонализированный сервис. Кластер 2 включает группу клиентов, обладающих высокой степенью реакции на ценовые стимулы. Для повышения их покупательской активности целесообразно использовать ограниченные по времени предложения, сезонные скидки и бонусные программы. Кластер 1 представлен преимущественно новыми или нерешительными клиентами с низким уровнем вовлеченности.

Кроме алгоритма k-means, который является одним из самых популярных методов кластеризации, существует и другой мощный подход – алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise). В отличие от k-means, который требует заранее задать количество кластеров и предполагает, что они имеют примерно одинаковую форму и размер, DBSCAN работает на основе плотности данных. Он способен выявлять кластеры произвольной формы, автоматически определять их количество и эффективно обрабатывать выбросы, исключая их из кластеров [5].

Алгоритм DBSCAN особенно полезен в ситуациях, когда данные содержат шум или когда кластеры не имеют четких границ. Он группирует точки, находящиеся в плотных областях, и игнорирует те, что расположены изолированно [6]. Это делает его незаменимым инструментом при анализе сложных, неструктурированных или географических данных, где традиционные методы, такие как k-means, могут давать искаженные результаты. Пример реализации алгоритма DBSCAN приведен на рис. 3.

```
# Выбор признаков для кластеризации
features = ['orders', 'avg_spend', 'discount_usage']
scaler = Standardscaler()
scaled_data = scaler.fit_transform(df[features])
# Применение DBSCAN
dbscan = DBSCAN (eps-1.2, min_samples=3)
df['cluster'] = dbscan.fit_predict(scaled_data)
```

Рис. 3. Пример реализация алгоритма DBSCAN

Алгоритм DBSCAN группирует клиентов по плотности данных и автоматически определяет количество кластеров. Получившиеся кластеры представлены на рис. 4. DBSCAN успешно справился с задачей кластеризации клиентов по признакам, выявив две плотные группы, соответствующие различным моделям поведения. Обнаружены два четко выраженных кластера: кластер 1 клиентов с низкой покупательской активностью и кластер 2 – с высокой. Кластеры сформированы на основе плотности, а не геометрической симметрии, что позволяет учитывать реальные особенности распределения данных.

Таким образом, можно сделать вывод, что кластеры, полученные с помощью k-means и DBSCAN на одном и том же массиве данных, не обязательно будут одинаковыми и чаще всего они различаются. Это связано с тем, что данные алгоритмы используют разные принципы кластеризации и по-разному интерпретируют структуру данных. Например, с помощью алгоритма k-means получены три кластера, а с помощью алгоритма DBSCAN – два.

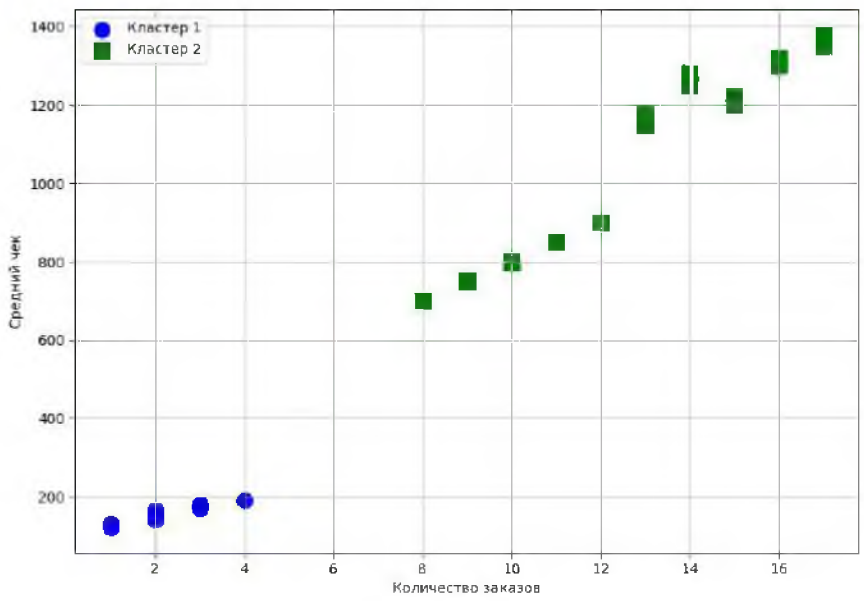

Рис. 4. Кластеризация с использованием алгоритма DBSCAN

Такие алгоритмы, как k-means и DBSCAN, являются мощным инструментом для принятия стратегических решений в бизнесе. K-means позволяет разбивать аудиторию на управляемые группы, что особенно полезно для маркетинга и планирования коммуникаций, тогда как DBSCAN способен выявлять скрытые ниши, плотные группы клиентов и аномальные паттерны поведения, которые не всегда видны при традиционном анализе. Зная к какому кластеру принадлежит клиент, компании могут точнее настраивать рекламу, скидки и персонализированные предложения, повышая эффективность взаимодействия. Сравнивая кластеры во времени, можно отслеживать динамику поведения клиентов и понимать, какие сегменты растут, а какие требуют дополнительного внимания.

Список литературы

1. E-commerce worldwide – statistics & facts. – URL: https://www.statista.com/topics/871/online-shopping/#topicOverview (date of access: 20.08.2025).

2. Миронов, А. М. Машинное обучение : учеб. пособие / А. М. Миронов. – М. : МГУ, 2023. – Ч. 1. – 215 с.

3. Кластеризация. Справочник по машинному обучению от Яндекс Образования. – URL: https://education.yandex.ru/handbook/ml/ article/klasterizaciya (дата обращения: 20.08.2025).

4. Кластеризация: k-means, DBSCAN и другие алгоритмы // Nerdit.ru. – URL: https://nerdit.ru/klastierizatsiia-k-means-dbscan-i-drughiie-alghoritmy/ (дата обращения: 28.06.2025).

5. Использование машинного обучения для сегментации клиентов. – URL: https://fastercapital.com/ru/content/Использование-машинного-обучения-для-сегментации-клиентов.html (дата обращения: 20.08.2025).

6. How does the DBSCAN algorithm work? // Medium. Analytics Vidhya. – URL: https://medium.com/analytics-vidhya/how-does-the-dbscan-algorithm-work-5a18098148d6 (date of access: 20.08.2025).