УДК 004

РАЗРАБОТКА МЕТОДИКИ СЕГМЕНТАЦИИ ПОЛЬЗОВАТЕЛЕЙ С ПОМОЩЬЮ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ И РАСШИРЕННОЙ АНАЛИТИКИ

Д. А. Клинов[1] [0000-0002-3623-9596], К. А. Григорян[2] [0000-0001-6470-1832]

[1, 2]Казанский (Приволжский) федеральный университет, ул. Кремлевская, 35, г. Казань, 420008

[1]daniil.klinov@bk.ru, [2]karigri@yandex.ru

Аннотация

Статья посвящена созданию эффективного решения по сегментации пользователей. Представлены анализ существующих сервисов сегментации пользователей и подходов к их сегментации (ABCDx сегментация, демографическая сегментация, сегментация на основании карты пути пользователя), а также анализ алгоритмов кластеризации (K-means, Mini-Batch K-means, DBSCAN, Agglomerative Clustering, Spectral Clustering). Исследование названных подходов нацелено на создание решения по сегментации, «гибкого» и адаптирующегося под каждую пользовательскую выборку. Также применены дисперсионный анализ (тест ANOVA) и разбор метрик кластеризации для оценки качества сегментации пользователей. С помощью указанных методов разработано эффективное решение по сегментации пользователей с использованием технологии расширенной аналитики и машинного обучения.

Ключевые слова: *Сегментация, кластеризация, дисперсионный анализ, машинное обучение, расширенная аналитика, тест ANOVA, продуктовая аналитика.*

ВВЕДЕНИЕ

В современном конкурентном мире крайне важно понимать поведение клиентов и классифицировать клиентов на основе их демографии и покупательского поведения. Это критический аспект сегментации клиентов, который позволяет

маркетологам лучше адаптировать свои маркетинговые усилия к различным подгруппам аудитории с точки зрения стратегий продвижения, маркетинга и разработки продуктов.

Сегментация пользователей — это процесс сегментирования группы лиц в соответствии с определенными характеристиками, чтобы максимально точно определить их ожидания и потребности. Исследования показывают, что сегментации клиентов помогают привести к тому, что компании тратят менее 20% рабочего времени своих сотрудников на развитие продукта для удовлетворения потребностей клиентов, приносящих более 80% общей выручки продукта [1].

Ведущие IT-компании разрабатывают свои внутренние алгоритмы сегментации клиентов. У небольших IT-компаний, работающих в B2C (коммерческие взаимоотношения между организацией и частными лицами), нет выделенных средств для создания эффективной сегментации. Исходя из этого, малому и среднему бизнесу приходится использовать базовые алгоритмы сегментации, которые не учитывают индивидуальную пользовательскую аналитику определенного продукта. Помимо этого, используемые открытые алгоритмы сегментации пользователей обладают рядом недочетов, которые можно избежать с помощью исследований в области расширенной аналитики [2].

Стремление разработчиков продуктов расширить критерии сегментации, чтобы включить интересы и предпочтения большого круга пользователей, приводит к проблеме устаревания используемого алгоритма сегментации. Данная статья направлена на изучение способов адаптации разрабатываемого алгоритма к изменению поведения пользователей продукта в результате взаимодействия с ним.

ИССЛЕДОВАНИЕ СУЩЕСТВУЮЩИХ СЕРВИСОВ ДЛЯ СЕГМЕНТАЦИИ ПОЛЬЗОВАТЕЛЕЙ

Рассмотрим особенности существующих решений.

- **BlueVenn –** https://www.bluevenn.com. Сервис предоставляет возможности аналитики данных, прогнозирования событий и сегментации пользователей. BlueVenn предоставляет API-интерфейс и удобную загрузку данных: настраиваемый механизм идентификационных данных позволяет обрабатывать, сопоставлять, объединять, избавляться от дубликатов в данных клиентов, дает

возможность работы с транзакциями клиентов, маркетинговыми каналами и источниками данных в режиме реального времени.

- **Commence Cloud CRM –** https://www.commence.com. Сервис предоставляет возможности по автоматизации демографической сегментации клиентов. Позволяет клиентам получать доступ к ряду заранее созданных сегментов, а пользователям – создавать новые демографические сегменты, объединяя данные о клиентах. Встроенная аналитика сегментации дает возможность проводить дальнейшие исследования и оценивать эффективность клиентских сегментов.
- **Qualtrics –** https://www.qualtrics.com**.** Сервис предоставляет возможности настройки исследований, создания целевых групп, анализа результатов исследования. Реализует сегментацию клиентов на единой платформе, что обеспечивает оперативный доступ к необходимым данным и сведениям о различных событиях.
- **Experian –** https://www.experian.com. Сервис предоставляет возможности настройки событий для сегментации клиентов. У пользователя есть возможность составлять и контролировать портфель наиболее прибыльных клиентов. Решения, предлагаемые этим ПО, в первую очередь направлены на идентификацию «портрета» клиента.
- **HubSpot –** https://www.hubspot.com. Продукт помогает работать с контактами клиентов, которые есть в базе данных пользователя. С помощью этого продукта пользователь может продумывать стратегии маркетинга, продаж и работы с клиентами.

Нами были изучены и проанализированы 5 сервисов сегментации пользователей, которые суммарно имеют на своих сайтах более 4 миллионов уникальных посещений в месяц. Все изученные сервисы отличаются высокой ценой и не используют в своих решениях технологии машинного обучения. Все сервисы используют фильтрацию данных и не располагают расширенной аналитикой для прогнозирования сегмента новых пользователей [3].

После детального анализа текущих решений в сегментации пользователей не удалось найти решения, которое бы использовало «гибкий» подход к выбору алгоритма сегментации пользователей в зависимости от набора атрибутов для

сегментации или индивидуальных особенностей продукта, пользователи которого сегментируются.

ОПИСАНИЕ АЛГОРИТМА СЕГМЕНТАЦИИ ПОЛЬЗОВАТЕЛЕЙ

Целью конечного алгоритма является выявление подгрупп пользователей (сегментов), отличающихся между собой покупательским потенциалом, активностью посещений и обращений в техническую поддержку, демографическими признаками или иными количественными и качественными метриками.

Проведем теперь анализ алгоритмов кластеризации.

K-Means – алгоритм на основе центроидов, в котором каждая точка данных размещается ровно в одном из K непересекающихся кластеров, выбранных до запуска алгоритма [4].

Mini-Batch K-Means – алгоритм использует небольшие случайные группы, так называемые «пакеты», размер которых установлен изначально, чтобы их можно было хранить в памяти, а затем при каждой итерации случайная выборка из набора данных собирается и используется для обновления кластеров [4].

DBSCAN – этому алгоритму требуются два параметра:

- *Eps*: если расстояние между двумя точками меньше или равно *eps*, то они считаются соседями. Если значение *eps* выбрано слишком маленьким, большая часть данных будет рассматриваться как выбросы. Если этот параметр выбран очень большим, то кластеры объединятся, и большинство точек данных будет в одних и тех же кластерах.

- *MinPts*: чем больше набор данных, тем должно быть выбрано большее значение *MinPts*. Как правило, минимальные *MinPts* могут быть получены из числа измерений D в наборе данных как MinPts>=D+1. Минимальное значение MinPts должно быть выбрано не меньшим 3 [7].

HAC (Hierarchical Agglomerative Clustering) – алгоритм, результатом реализации которого является древовидное представление объектов, называемое «дендрограммой», которая показывает прогрессивную группировку данных. Этот алгоритм кластеризации не требует предварительного указания количества кластеров. Алгоритмы «снизу вверх» сначала обрабатывают данные как отдельный кластер, а затем последовательно объединяют пары кластеров, пока все кластеры не будут объединены в один кластер, содержащий все данные [7].

Метод спектральной кластеризации – этот алгоритм использует информацию из собственных значений (спектра) специальных матриц, построенных из графика или набора данных. Он рассматривает каждую точку данных как узел графа и, таким образом, преобразует задачу кластеризации в задачу разделения графа. Метод спектральной кластеризации не делает сильных предположений о статистике кластеров – в отличие от алгоритма K-средних, который предполагает, что точки, назначенные кластеру, имеют сферическую форму относительно центра кластера. В таких случаях спектральная кластеризация помогает создавать более точные кластеры [4–7].

Этапы сегментации пользователей сводятся к последовательному выполнению следующих шагов:

1. Загрузка пользовательских данных;
2. Выбор подхода к сегментации пользователей: ABCDx, Demographics, Journey map;
3. Реализация метода главных компонент для определения атрибутов кластеризации;
4. Применение алгоритмов кластеризации к пользовательским данным: K-means, Mini-Batch K-means, DBSCAN, Agglomerative Clustering, Spectral Clustering;
5. Анализ эффективности метрик алгоритмов кластеризации;
6. Применение теста ANOVA на тех же пользовательских данных для оценки количественных показателей эффективности алгоритма кластеризации;
7. Определение наиболее эффективного алгоритма кластеризации на данных, загруженных пользователем;
8. Финальная сегментация пользователей по определенному алгоритму кластеризации.

АНАЛИЗ ЭФФЕКТИВНОСТИ АЛГОРИТМА КЛАСТЕРИЗАЦИИ

Метрики качества кластеризации

Прежде чем анализировать качество кластеризации, нами определен термин *«эталонный кластер». Эталонные* кластеры существуют на исследуемом

множестве независимо от алгоритма кластеризации. Кластеризация на эталонные кластеры – это лучший результат работы алгоритма кластеризации среди всех возможных результатов, принимая во внимание, что может не существовать алгоритма, обеспечивающего кластеризацию на эталонные кластеры.

Анализ метрик кластеризации

1. Однородность кластеров. Качество кластеризации ухудшается, если происходит объединение двух эталонных кластеров в один (Рис. 1).

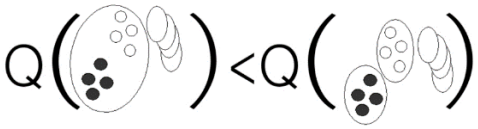

Рис. 1. Качество кластеризации в зависимости от однородности кластеров.

2. Полнота кластеров. Качество кластеризации ухудшается, если происходит разделение эталонного кластера на два других кластера (Рис. 2).

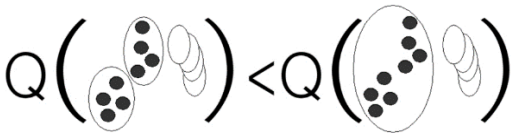

Рис. 2. Качество кластеризации в зависимости от полноты кластеров.

3. Чистота кластеров. Пусть на множестве есть эталонный кластер и несколько нерелевантных элементов, каждый из которых представляет эталонный кластер. Качество кластеризации увеличивается, если эталонный кластер выделяется в отдельный кластер без добавления других элементов (Рис. 3).

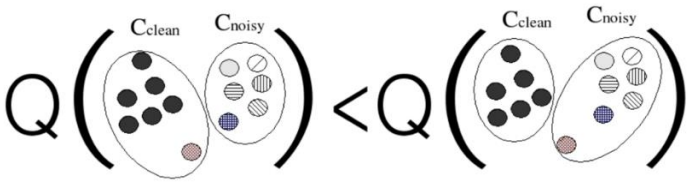

Рис. 3. Качество кластеризации в зависимости от чистоты кластеров.

4. Количество и размер кластеров. Качество кластеризации ухудшается, если отсутствует большое число небольших эталонных кластеров, но при этом присутствует один крупный эталонный кластер [8] (Рис. 4).

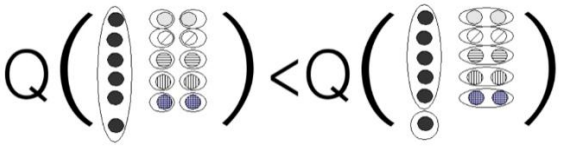

Рис. 4. Качество кластеризации в зависимости размера и количества кластеров.

Тест ANOVA

Правила, по которым применяется тест ANOVA для анализа качества кластеризации пользовательских данных, состоят в следующем:

- Исследуемой выборкой является база данных пользователей;
- Подгруппами являются сегменты пользователей, определенные по результатам кластеризации;
- Качественной характеристикой является сегмент пользователя;
- Количественными характеристиками являются генерируемая пользователем прибыль, сессии пользователя, обращения пользователя в техническую поддержку.

Метрикой качества кластеризации является статистическая значимость отличий между сформированными подгруппами (кластерами). Чем большую статистическую значимость имеют зависимости количественных характеристик от подгрупп, тем качественней считается алгоритм кластеризации.

На основании метрик кластеризации и теста ANOVA определяется эффективность алгоритма кластеризации [9].

ЗАКЛЮЧЕНИЕ

В результате проведенного исследования были проанализированы существующие сервисы сегментации пользователей. Было разработано решение с использованием различных алгоритмов кластеризации для эффективной сегментации пользовательских данных.

Существующие подходы к сегментации пользователей не являются «гибкими», они не адаптируются под определенные пользовательские данные. Решение, разобранное нами, предполагает анализ качества нескольких алгоритмов кластеризации с помощью метрик кластеризации и теста ANOVA и последующее использование эффективного алгоритма кластеризации.

Продолжить развитие данного исследования можно в сторону развития базы алгоритмов кластеризации, подходов к сегментации, улучшения оценки качества кластеризации и эффективности сегментации.

СПИСОК ЛИТЕРАТУРЫ

1. *Чурин В.В.* Роль маркетинговых исследований в проектной деятельности: Учебно-методическое пособие // Московский автомобильно-дорожный государственный технический университет (МАДИ). 2019. С. 1–111.

2. *An J., Kwak H., Jung S., Salminen J., Jansen B.* Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data // Social Network Analysis and Mining. 2018. P. 1–19.

3. *Старкова Н В.* Кластеризация стран Европы по демографическим признакам // Молодой ученый. 2016. № 9 (113). С. 418–426.
URL: https://moluch.ru/archive/113/28811/ (дата обращения: 06.06.2022)

4. *Черезов Д.С.* Обзор основных методов классификации и кластеризации данных // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2009. №2. С. 23–27.
URL: https://rucont.ru/efd/519732 (дата обращения: 06.06.2022)

5. *Jagabathula S., Rusmevichientong P., Venkataraman A., Zhao X.* Estimating Large-Scale Tree Logit Models // NYU Stern School of Business, 2022.

6. *Amigó E., Gonzalo J., Artiles J.* A comparison of extrinsic clustering evaluation metrics based on formal constraints // Information Retrieval volume. 2009. No. 12. P. 461–486.

7. *Топалович Н*. Алгоритмы кластеризации в машинном обучении // Молодой ученый. 2020. № 52 (342). С. 47–49.
URL: https://moluch.ru/archive/342/77003/ (дата обращения: 06.06.2022)

8. *Rodriguez M.Z., Comin C.H., Casanova D., Bruno O.M., Amancio D.R., Costa L.F., Rodrigues F.A.* Clustering algorithms: A comparative approach // PLoS One. 2019. No. 14. P. 15–30.

9. *Байков И.И.* Метод ансамблирования алгоритмов кластеризации для решения задачи совместной кластеризации // Сенсорные системы. 2021. Т. 35. № 1. С. 43–49.

DEVELOPMENT OF A METHOD FOR USER SEGMENTATION USING CLUSTERING ALGORITHMS AND ADVANCED ANALYTICS

D. A. Klinov[1] [0000-0002-3623-9596], K. A. Grigorian[2] [0000-0001-6470-1832]

[1, 2]Kazan (Volga Region) Federal University, 35 Kremlevskaya str., Kazan, 420008

[1]daniil.klinov@delion.ru, [2]karigri@yandex.ru

Abstract

The article is devoted to the creation of an effective solution for user segmentation. The article presents an analysis of existing user segmentation services, an analysis of approaches to user segmentation (ABCDx segmentation, demographic segmentation, segmentation based on a user journey map), an analysis of clustering algorithms (K-means, Mini-Batch K-means, DBSCAN, Agglomerative Clustering, Spectral Clustering). The study of these areas is aimed at creating a “flexible” segmentation solution that adapts to each user sample. Dispersion analysis (ANOVA test), analysis of clustering metrics is also used to assess the quality of user segmentation. With the help of these areas, an effective solution for user segmentation has been developed using advanced analytics and machine learning technology.

Keywords: *Segmentation, clustering, analysis of variance, machine learning, advanced analytics, ANOVA test, product analytics.*

REFERENCES

1. *Churin V.V.* Rol' marketingovyh issledovanij v proektnoj dejatel'nosti: Uchebno-metodicheskoe posobie // Moskovskij avtomobil'no-dorozhnyj gosudarstvennyj tehnicheskij universitet (MADI). 2019. S. 1–111.

2. *An J., Kwak H., Jung S., Salminen J., Jansen B.* Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data // Social Network Analysis and Mining. 2018. P. 1–19.

3. *Starkova N V.* Klasterizacija stran Evropy po demograficheskim priznakam // Molodoj uchenyj. 2016. № 9 (113). S. 418–426.
URL: https://moluch.ru/archive/113/28811/ (date of the application: 06.06.2022)

4. *Cherezov D.S.* Obzor osnovnyh metodov klassifikacii i klasteri-zacii dannyh // Vestnik Voronezhskogo gosudarstvennogo universiteta. Serija: Sistemnyj analiz i informacionnye tehnologii. 2009. №2. S. 23–27. URL: https://rucont.ru/efd/519732 (date of the application: 06.06.2022)

5. *Jagabathula S., Rusmevichientong P., Venkataraman A., Zhao X.* Estimating Large-Scale Tree Logit Models // NYU Stern School of Business, 2022.

6. *Amigó E., Gonzalo J., Artiles J.* A comparison of extrinsic clustering evaluation metrics based on formal constraints // Information Retrieval volume. 2009. No. 12. P. 461–486.

7. *Topalovich N.* Algoritmy klasterizacii v mashinnom obuchenii // Molodoj uchenyj. 2020. № 52 (342). S. 47–49.
URL: https://moluch.ru/archive/342/77003/ (date of the application: 06.06.2022)

8. *Rodriguez M.Z., Comin C.H., Casanova D., Bruno O.M., Amancio D.R., Costa L.F., Rodrigues F.A.* Clustering algorithms: A comparative approach // PLoS One. 2019. No. 14. P. 15–30.

9. *Bajkov I.I.* Metod ansamblirovanija algoritmov klasterizacii dlja reshenija zadachi sovmestnoj klasterizacii // Sensornye sistemy. 2021. T. 35. № 1. S. 43–49.

СВЕДЕНИЯ ОБ АВТОРАХ

КЛИНОВ Даниил Андреевич[1] – магистр Института информационных технологий и интеллектуальных систем по направлению «Программная инженерия», изучает продуктовую аналитику.

KLINOV Daniil Andreevic[1]– the magister of Institute of Information Technologies and Intelligent Systems in the direction “Software engineering”, middle of product analytics.

Email: daniil.klinov@bk.ru
ORCID: 0000-0002-3623-9596

ГРИГОРЯН Карен Альбертович[2] – кандидат экономических наук, доцент, Казанский (Приволжский) федеральный университет, г. Казань.

GRIGORIAN Karen Albertovich[2] – candidate of Economics, Associate Professor, Kazan (Volga region) Federal University, Kazan.

Email: karigri@yandex.ru
ORCID: 0000-0001-6470-1832

Материал поступил в редакцию 31 мая 2022 года