УДК 004.85:366.14

152. ПРИМЕНЕНИЕ АЛГОРИТМОВ КЛАССИФИКАЦИИ И КЛАСТЕРИЗАЦИИ В DATA MINING ДЛЯ СЕГМЕНТАЦИИ ПОЛЬЗОВАТЕЛЕЙ IT-ПРОДУКТОВ

Мягков А.А.студент гр. 578105, Шидловский К. Д. студент гр. 372301

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Петрович Ю.Ю. - старший преподаватель кафедры ЭИ, магистр

Аннотация. В данной работе рассматриваются теоретические аспекты применения технологий интеллектуального анализа данных (Data Mining) в задачах сегментации аудитории цифровых сервисов. Основное внимание уделено методологическим различиям между алгоритмами кластеризации и классификации, а также их синергии при формировании стратегий персонализации в IT-индустрии. Описаны ключевые этапы обработки данных и принципы выделения пользовательских паттернов, позволяющие повысить эффективность удержания и вовлеченности клиентов.

Ключевые слова. Data Mining, сегментация пользователей, кластеризация, классификация, машинное обучение, пользовательское поведение, персонализация.

В условиях глобальной цифровизации и экспоненциального роста объемов генерируемой информации классические методы маркетингового анализа и статистической отчетности перестают удовлетворять потребностям динамично развивающихся IT-компаний. Современный IT-продукт — будь то мобильное приложение, облачная платформа или социальная сеть — ежедневно накапливает массивы данных о действиях пользователей. Эти данные представляют собой сложный цифровой след, анализ которого методами Data Mining позволяет перейти от простого наблюдения к проактивному моделированию пользовательского опыта. Сегментация пользователей является фундаментальной задачей, решение которой позволяет разделить неоднородную аудиторию на группы с идентичными паттернами поведения. Это необходимо для решения широкого спектра задач: от снижения уровня оттока до максимизации среднего дохода на пользователя. Научный интерес к данной теме обусловлен поиском наиболее устойчивых и интерпретируемых алгоритмов, способных эффективно работать в условиях высокой размерности данных и информационного шума. При этом на практике особую сложность представляет выбор адекватной меры расстояния между объектами: евклидова метрика часто оказывается малоэффективной в пространствах высокой размерности из-за эффекта «проклятия размерности», что делает предпочтительным использование мер на основе косинусного сходства или корреляции Манхэттена.

Критически важным этапом, предшествующим применению алгоритмов, является проектирование признаков, поскольку качество работы любой модели напрямую ограничено качеством входных данных. В контексте цифровых продуктов признаковое пространство обычно формируется из нескольких категорий, включая поведенческие показатели, такие как частота использования функций и длительность сессий, а также транзакционные и технические характеристики. Методологически значимым процессом здесь выступает нормализация и стандартизация данных, так как многие алгоритмы (особенно методы, основанные на вычислении расстояний, и метод главных компонент) чувствительны к масштабу переменных. На практике применяются два основных подхода: min-max нормализация, приводящая значения к интервалу [0,1], и z-стандартизация, центрирующая данные относительно нулевого среднего с единичным дисперсионным разбросом; выбор между ними определяется наличием выраженных выбросов и требованием сохранения формы исходного распределения. Кроме того, в современных исследованиях уделяется внимание методам снижения размерности, которые позволяют сохранить наиболее значимую информацию при уменьшении количества переменных, что снижает вычислительную сложность и вероятность переобучения моделей. Конкретными инструментами здесь выступают не только линейный метод главных компонент, но и нелинейные методы, такие как t-SNE и UMAP, которые эффективно визуализируют кластерную структуру, однако требуют осторожной интерпретации из-за стохастичности результата.

После завершения этапа подготовки данных на первый план выходит методология кластеризации, которая классифицируется как обучение без учителя. Основная цель данного подхода заключается в группировке объектов таким образом, чтобы внутригрупповое сходство было максимальным, а межгрупповое — минимальным. В теории анализа данных выделяют несколько ключевых подходов к кластеризации пользовательской базы. Наиболее распространенным является метод k-средних, основанный на итеративном поиске центроидов групп. Он ценится за вычислительную эффективность и простоту интерпретации, хотя и требует предварительного обоснования количества сегментов через использование вспомогательных техник, таких как метод «локтя» или индекс силуэта, причем последний позволяет не только выбрать оптимальное число кластеров, но и оценить степень компактности каждого выделенного сегмента. Важной практической проблемой метода k-средних является чувствительность к начальному выбору центроидов, что частично решается применением инициализации k-means++ или многократным запуском алгоритма с последующим выбором результата с наименьшей суммой квадратов внутрикластерных расстояний. В свою очередь, иерархические методы позволяют строить структуру вложенных кластеров, что полезно при анализе сложных зависимостей, когда внутри крупных сегментов необходимо выделить узкие подгруппы по интересам. При этом выбор меры связи между кластерами — одиночной, полной или средней — существенно влияет на форму получаемых дендрограмм, и для поведенческих данных чаще применяется метод Уорда, минимизирующий внутрикластерную дисперсию. Еще одним важным направлением являются плотностные методы, которые определяют кластеры как области высокой концентрации данных. Это позволяет выделять сегменты произвольной геометрической формы и эффективно отсеивать аномальных пользователей, чье поведение может исказить общие результаты анализа. Наиболее репрезентативным представителем данной группы выступает алгоритм DBSCAN, требующий задания двух параметров: радиуса окрестности и минимального числа точек в окрестности; их подбор часто осуществляется с помощью анализа k-графиков расстояний, что является эмпирической, но технически обоснованной процедурой.

Параллельно с методами поиска скрытых структур активно развиваются алгоритмы классификации, представляющие собой обучение с учителем. Они применяются для оперативного распределения новых пользователей по уже известным категориям. Основными инструментами здесь выступают деревья решений, позволяющие визуализировать логику сегментации в виде последовательных условий, что делает модель прозрачной для интерпретации специалистами. Однако классические деревья решений склонны к переобучению при наличии шумовых признаков, что ограничивает их применение на сырых поведенческих данных без предварительной фильтрации. Более сложные ансамблевые методы, такие как случайный лес, позволяют значительно повысить точность классификации и устойчивость системы к случайным отклонениям в данных за счет агрегирования большого количества слабо коррелированных деревьев. В задачах с крайне высокой размерностью признаков, например, при анализе текстовых логов или сложных интерфейсных взаимодействий, применяются методы опорных векторов и глубокие нейронные сети. Для последних ключевой проблемой становится обеспечение интерпретируемости, что частично решается методами пост-хок объяснений, такими как LIME или SHAP, позволяющими аппроксимировать решение нейронной сети локально-интерпретируемой моделью в окрестности конкретного пользовательского экземпляра.

Наибольшую научную ценность в данном контексте представляет гибридный подход, при котором кластеризация и классификация применяются последовательно. На первом этапе исследователь формирует архитектуру сегментов, выявляя устойчивые типы пользователей на основе исторического массива данных. Это позволяет не навязывать продукту субъективные категории, а обнаружить реальные поведенческие паттерны. На втором этапе полученные метки используются для обучения классификатора, который в реальном времени начинает определять принадлежность нового клиента к конкретной группе уже после первых минут его активности. Такая синергия обеспечивает возможность мгновенной адаптации продукта, что является высшим уровнем развития персонализации в современной IT-индустрии. При этом важно, чтобы классификатор на этапе продуктивного использования периодически переобучался на свежих данных, поскольку распределение поведенческих паттернов может дрейфовать — явление, известное как концептуальный дрейф (concept drift). Для его обнаружения применяются методы мониторинга ошибки классификации на контрольной выборке или статистические тесты на однородность распределений, например, на основе двустороннего критерия Колмогорова-Смирнова.

Несмотря на высокую эффективность описанных методов, их внедрение сопряжено с рядом методологических проблем, одной из которых является изменчивость поведения пользователей во времени. Это требует разработки систем непрерывного обучения моделей для сохранения их

актуальности. Кроме того, в академической среде ведется активная дискуссия об этике использования алгоритмов интеллектуального анализа. Глубокая сегментация может приводить к созданию «информационных пузырей» или неосознанной дискриминации определенных групп. Конкретные этические риски включают возможность косвенной дискриминации по косвенным признакам (например, использование геолокации как прокси для расовой или экономической принадлежности) и эффект самосбывающегося прогноза, когда алгоритмически назначенная категория влияет на пользовательский опыт, искусственно закрепляя пользователя в этой категории. Следовательно, при проектировании систем сегментации необходимо соблюдать баланс между эффективностью алгоритмов и принципами прозрачности и справедливости принимаемых ими решений. Практическим инструментом здесь выступает метрика демографического паритета, позволяющая количественно оценить, насколько равномерно представлены различные социальные группы в выделенных сегментах.

Подводя итог проведенному анализу, можно утверждать, что использование методов Data Mining позволяет трансформировать разрозненные данные активности в стратегический актив компании. Теоретический обзор подтверждает, что выбор конкретных инструментов должен быть обусловлен целями исследования, объемом данных и требуемым уровнем интерпретируемости результатов. Будущее направление исследований в данной области связано с интеграцией методов анализа графов и глубокого обучения, что позволит еще более точно моделировать сложные социальные и поведенческие связи внутри современных цифровых экосистем. В частности, графовые нейронные сети (GNN) открывают возможность учитывать не только индивидуальное поведение пользователя, но и структуру его социальных связей и влияние соседей по графу взаимодействий, что особенно актуально для социальных сетей и многопользовательских платформ.

Список использованных источников:

1. *Han, J., Kamber, M., Pei, J. Data Mining: Concepts and Techniques / J. Han, M. Kamber, J. Pei. − 3rd ed. − Waltham : Morgan Kaufmann, 2011. − 744 p.*
2. *Provost, F., Fawcett, T. Data Science for Business / F. Provost, T. Fawcett. − Sebastopol : O'Reilly Media, 2013. − 414 p.*
3. *Вьюгин, В. В. Математические основы теории машинного обучения и прогнозирования / В. В. Вьюгин. − Москва : МЦНМО, 2013. − 387 с.*
4. *Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. Friedman. − 2nd ed. − New York : Springer, 2009. − 745 p.*
5. *Pedregosa, F. Scikit-learn: Machine Learning in Python / F. Pedregosa [et al.] // Journal of Machine Learning Research. − 2011. − Vol. 12. − P. 2825−2830.*
6. *Флах, П. Машинное обучение. Наука и искусство построения алгоритмов / П. Флах. − Москва : ДМК Пресс, 2015.*

UDC 004.85:366.14

INVESTIGATION OF THE ACOUSTIC PROPERTIES OF ROOMS USING THE REVERBERATION TIME MEASUREMENT METHOD

Miahkou A.A. students of gp.578105, Shidlovsky K. D. Student Group 372301

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Petrovich Y.Y. - Senior lecturer at the Department of EI , Master's degree

Annotation. This paper examines the theoretical aspects of the application of data mining technologies in the tasks of audience segmentation of digital services. The main focus is on the methodological differences between clustering and classification algorithms, as well as their synergy in shaping personalization strategies in the IT industry. The key stages of data processing and the principles of identifying user patterns are described, which make it possible to increase the effectiveness of customer retention and engagement.

Keywords. Data Mining, сегментация пользователей, кластеризация, классификация, машинное обучение, пользовательское поведение, персонализация.