УДК 004.4, 004.6, 004.942, 004.021: 519.6, 616.12

https://doi.org/10.17721/1812-5409.2022/4.8

Н. О. Михайлов, аспірант

N. O. Mykhailov, PhD student

Класифікація користувачів на онлайн платформах методами машинного навчання

Classifications of users on online platforms using machine learning techniques

Київський національний університет імені Тараса Шевченка, Україна, 03022, м.Київ, просп. Академіка Глушкова, 4Д, 03022
e-mail: nikmikhailov13@gmail.com

Taras Shevchenko National University of Kyiv, 4D, Academician Glushkov ave., Kyiv, Ukraine, 03022
e-mail: nikmikhailov13@gmail.com

Онлайн-платформи стали невід'ємною частиною нашого життя, а кількість інтернет користувачів зростає з кожним днем. Від платформ соціальних медіа до веб-сайтів електронної комерції, цими платформами користуються мільйони людей у всьому світі. З такою великою базою юзерів цим платформам важливо класифікувати своїх користувачів на основі їх поведінки, уподобань та інтересів. У цій статті досліджується, як машинне навчання можна використовувати для класифікації користувачів на онлайн-платформах.

Класифікуючи користувачів, їх поділяють на різні категорії на основі їхніх характеристик. Аналізуючи поведінку та вподобання користувачів, онлайн-платформи можуть персоналізувати свої послуги та забезпечувати кращий досвід користувачів. Методи машинного навчання можуть допомогти онлайн-платформам автоматизувати процес класифікації та зменшити людські зусилля. У цій статі буде детально розглянута поведінкова класифікація користувачів на онлайн-платформах.

Ключові слова: класифікація користувачів, аналіз користувачів, кластеризація, правила асоціацій, нейронні мережі.

Online platforms have become an integral part of our lives, and the number of users is increasing by the day. From social media platforms to e-commerce websites, these platforms are used by millions of people around the world. With such a large user base, it is essential for these platforms to classify their users based on their behavior, preferences, and interests. This paper explores how machine learning can be used to classify users on online platforms.

When classifying users, they are divided into different categories based on their characteristics. By analyzing user behavior and preferences, online platforms can personalize their services and provide a better user experience. Machine learning techniques can help online platforms automate the classification process and reduce human effort. In this article, the behavioral classification of users on online platforms will be discussed in detail.

Keywords: user classification, user analysis, cluster analysis, association rules, neural networks.

© Нікіта Олегович Михайлов, 2022

Вступ

Зі збільшенням кількості користувачів на онлайн платформах стало надзвичайно важливо класифікувати користувачів на основі їх поведінки, уподобань та інтересів. Алгоритми штучного інтелекту довели свою ефективність у автоматизації процесу класифікації користувачів і наданні користувачам персоналізованого досвіду.

На сьогоднішній день інсує велика калькість алгоритмів штучного інтелекту (ШІ), які використовуються для класифікації користувачів на онлайн-платформах, включаючи демографічну класифікацію, класифікацію поведінки та класифікацію настроїю. У цій статі в центрі уваги буде поведінкова класифікація.

Поведінкова класифікація

Поведінкова класифікація — це процес класифікації користувачів на основі їх поведінки на онлайн-платформі. Ця класифікація базується на активності користувачів, наприклад відвідуваних ними сторінках, продуктах, які вони купують, публікаціях, якими вони діляться, і часу, який вони проводять на платформі.

Онлайн-платформи використовують поведінкову класифікацію, щоб персоналізувати свої послуги для користувачів, надавати кращі рекомендації та покращувати залучення користувачів.

Алгоритми штучного інтелекту, такі як кластеризація, правила асоціації та нейронні мережі, зазвичай використовуються для даної класифікації. Ці алгоритми аналізують дані користувачів і створюють моделі, які можуть передбачати поведінку та вподобання користувачів на основі їхньої минулої активності на платформі.

Кластеризація

Кластерний аналіз або кластеризація — це групування набору об’єктів таким чином, щоб об’єкти в одній групі (які називаються кластерами) були більш схожими один на одного (у певному сенсі), ніж об’єкти в інших групах (кластерах). Це головне завдання дослідницького аналізу даних і загальна техніка статистичного аналізу даних, яка використовується в багатьох галузях, включаючи розпізнавання образів, аналіз зображень, пошук інформації, біоінформатику, стиснення даних, комп’ютерну графіку та машинне навчання [1].

Сам по собі кластерний аналіз - це не конкретний алгоритм, а загальне завдання, яке необхідно вирішити. Його можна вирішити різними алгоритмами, які суттєво відрізняються у розумінні того, що таке кластер і як його ефективно знаходити. Загальні поняття кластерів включають групи з малими відстанями між членами кластерів, щільні області простору даних, інтервали або певні статистичні розподіли. Тому кластеризація може бути сформульована як багатоцільова задача оптимізації [2]. Відповідний алгоритм кластеризації та налаштування параметрів (включаючи такі параметри, як функція відстані для використання, поріг щільності або кількість очікуваних кластерів) залежать від конкретного набору даних і передбачуваного використання результатів. Кластерний аналіз як такий не є автоматичним завданням, а повторюваним процесом виявлення знань або інтерактивної багатоцільової оптимізації, яка передбачає спроби та помилки. Часто необхідно змінювати попередню обробку даних і параметри моделі, поки результат не матиме бажаних властивостей.

На додаток до терміну кластеризація існує ряд термінів зі схожими значеннями, включаючи автоматичну класифікацію, числову таксономію, ботріологію, типологічний аналіз і виявлення спільноти. Незначні відмінності часто полягають у використанні результатів: у той час як при інтелектуальному аналізі даних цікавлять отримані групи, при автоматичній класифікації цікавить результуюча відмінність [3].

Кластерний аналіз був розроблений в антропології Драйвером і Кробером в 1932 році і введений в психологію Джозефом Зубіним в 1938 році і Робертом Трайоном в 1939 році. Його прославив Кеттелл, який використовував його для класифікації рис у психології особистості з 1943 року [4].

Правила асоціації

Навчання правилам асоціації — це метод машинного навчання на основі правил для виявлення цікавих зв'язків між змінними у великих базах даних. Він використовується для визначення сильних правил, виявлених у базах даних, за допомогою певних показників цікавості. У будь-якій транзакції з великою кількістю елементів правила асоціації призначені для виявлення правил, які визначають, як і чому певні елементи пов'язані [5].

Базуючись на концепції сильних правил, Ракеш Агравал, Томаш Імелінскі та Арун Свамі запровадили правила асоціації, щоб виявити закономірності між продуктами у даних великих транзакцій, записаних касовими системами в супермаркетах. Наприклад, правило «цибуля, картопля — гамбургери», знайдене в даних про продажі в супермаркеті, означатиме, що клієнт, який купує разом цибулю та картоплю, швидше за все, купить гамбургери. Така інформація може бути використана як основа для прийняття рішень щодо маркетингових заходів, таких як ціни на рекламу або розміщення продуктів [6].

На додаток до прикладу аналізу кошика для покупок, наведеного вище, правила асоціації тепер використовуються в багатьох областях застосування, таких як оцінка використання веб-сайтів, безперервне виробництво та біоінформатика. На відміну від інтелектуального аналізу послідовностей, навчання правил асоціації зазвичай не враховує порядок елементів у транзакції чи між транзакціями [7].

Нейронні мережі

Нейронна мережа — це мережа або схема, що є або біологічною нейронною мережею, що складається з біологічних нейронів, або штучною нейронною мережею, яка використовується для вирішення проблем штучного інтелекту. Зв'язки біологічного нейрона моделюються в штучних нейронних мережах як ваги між вузлами. Позитивна вага являє собою збудливий зв'язок, тоді як негативні значення означають гальмівний зв'язок. Усі вхідні дані модифікуються за вагою та підсумовуються [8]. Ця діяльність називається лінійною комбінацією. Нарешті, функція активації контролює амплітуду вихідного сигналу. Наприклад, прийнятний діапазон для виведення зазвичай становить від 0 до 1 або може бути від -1 до 1.

Ці штучні мережі можна використовувати для прогнозного моделювання, адаптивного керування та додатків, де їх можна навчати на наборі даних. Мережі можуть самонавчатися і робити висновки зі складного та, здавалося б, непов'язаного набору інформації. [9]

Кластеризація методом к-середніх

Кластеризація методом к-середніх — популярний метод кластеризації, впорядкування множини об'єктів в порівняно однорідні групи. Винайдений в 1950-х роках математиком Гуґо Штайнгаузом і майже одночасно Стюартом Ллойдом.

Мета методу — розділити n спостережень на k кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера.

У цій статі буде продемонтрована робота алгоритму к-середніх на основі синтетичного датасету що складаєється з тисячі умовних користувачів онлайн платформи, які мають дві довільні характеристики. Кількість ознак користувачів була обрана для поліпшення сприйняття алгоритму та полегшення зображення користувачів на системі координат. Датасет було сгенеровано використовуючи бібліотеку NumPy мови програмування Python [10].

NumPy — це потужна бібліотека на Python, яка використовується для чисельних обчислень та надає набір потужних інструментів і структур даних для роботи з масивами і матрицями. NumPy використовується для виконання складних математичних операцій, таких як лінійна алгебра, перетворення Фур'є та генерація випадкових чисел. Ця бібліотека надає велику кількість математичних функцій і операторів, які працюють

з багатовимірними масивами, що робить її цінним інструментом для наукових обчислень [11].

Розміщення датасету на графіку виглядає наступним чином (див. рис. 1).

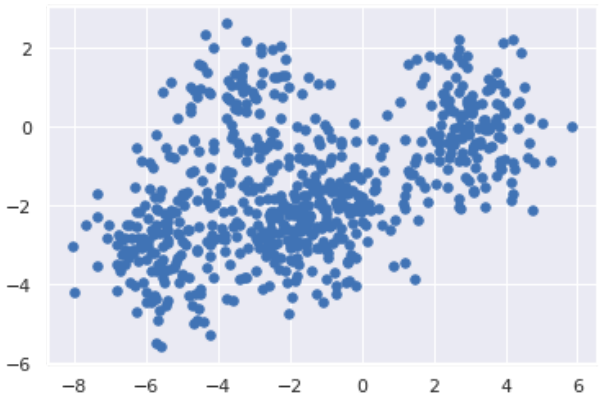

Рис. 1. Датасет користувачів онлайн платформи

K-Means — це дуже простий алгоритм, який об'єднує дані в K кластерів. Припустиму, що ми маємо вхідні дані X1, X2, X3 … Xn і значення K:

1. Виберемо K випадкових точок як центри кластерів, які називаються центроїдами
2. Примвоїмо кожен Xi найближчому кластеру, обчисливши його відстань до кожного центроїда
3. Знайдемо новий центр кластера, взявши середнє значення призначених точок
4. Повторимо кроки 2 і 3, доки центри нових кластерів більше не будуть змінюватись.

На першій ітерації алгоритму центроїди будуть знаходитися в наступних місцях (див. рис. 2) :

Рис. 3. Розміщення центроїдів на першій ітерації

Фінальний результат роботи алгоритму можно побачити на рис. 3.

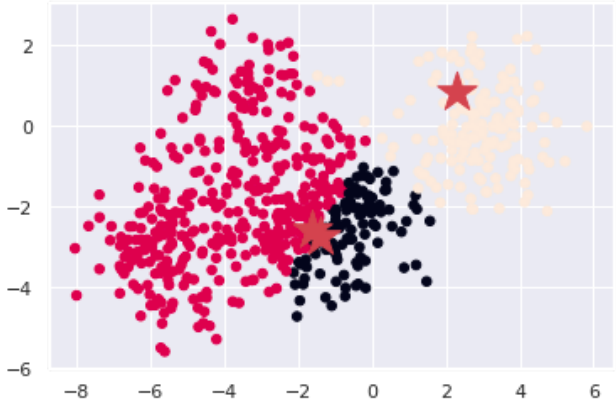

Рис. 3. Результат роботи алгоритму

Для перевірки коректонсті роботи алгоритму і правильності визначення кількості кластерів K ми можемо скористатися методом "Elbow". Для цього нам потрібно обчислити середню суму квадратів відстані між точками та центроїдом всередині кластера. З графіку на рис. 4 видно що K=3 є оптимальним значенням для даного датасету.

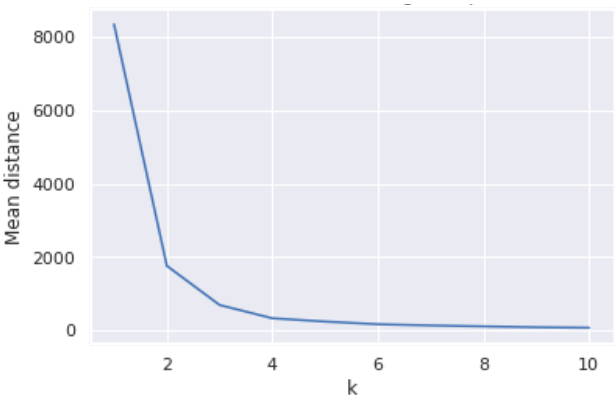

Рис. 4. Графік ортимальності вибору K за методом "Elbow"

Висновки

Використовуючи можливості розглянутих алгоритмів машинного навчання, онлайн сервіси та платформи можуть надавати персоналізовані послуги, покращувати взаємодію з клієнтами, виявляти і попереджувати шахрайство та спам

тим самим стаючи впевненими лідерами у своєму сегменті.

Розглянутий алгоритм K-Means — це потужний алгоритм кластеризації, який можна використовувати для класифікації користувачів на онлайн-платформах.

Однією з ключових переваг алгоритму K-Means є його здатність автоматично групувати користувачів у кластери на основі їхньої поведінки чи вподобань. Це означає, що його можна використовувати для ідентифікації груп користувачів, які мають подібні характеристики, наприклад історію веб-перегляду, поведінку при купівлі або демографічну інформацію.

Ще одна перевага використання алгоритму K-Means для класифікації користувачів полягає в тому, що він може допомогти покращити взаємодію з користувачем на онлайн-платформах. Групуючи користувачів у кластери на основі їхніх уподобань, платформи можуть надавати персоналізовані рекомендації та контент користувачам.

Загалом, алгоритм K-Means є важливим інструментом для онлайн-платформ, які прагнуть краще зрозуміти своїх користувачів і покращити взаємодію з ними.

Продуктивність алгоритму K-Means, нейронних мереж і правил асоціації може відрізнятися залежно від характеру даних, їх розміру і складності проблеми класифікації.

Алгоритм K-Means є обчислювально ефективним, особливо при роботі з великими наборами даних, і його легко інтерпретувати та візуалізувати результати. Однак він обмежений лінійними межами і не може обробляти складні нелінійні розподіли даних.

З іншого боку, нейронні мережі — це потужні алгоритми машинного навчання, які можуть вивчати складні патерни в даних і добре виконувати широкий спектр завдань класифікації, включаючи класифікацію користувачів. Однак нейронні мережі є дорогими в обчислювальному плані, потребують багато даних для навчання та можуть бути складними для інтерпретації.

Правила асоціації — це ще один тип алгоритму машинного навчання, який можна використовувати для класифікації користувачів. Правила асоціації працюють шляхом виявлення частих шаблонів у даних і створення правил, які фіксують зв’язки між різними змінними. Правила асоціації прості для інтерпретації та можуть добре працювати для наборів даних із великою кількістю змінних. Однак вони можуть бути чутливими до шуму та аномалій у даних і можуть показувати некоректні результати при роботі з даними великої розмірності.

Отже, вибір алгоритму для класифікації користувачів на онлайн-платформах залежить від конкретних характеристик набору даних і проблеми класифікації.

Список використаних джерел

1. Driver and Kroeber. Quantitative Expression of Cultural Relationships // University of California Publications in American Archaeology and Ethnology. Berkeley, CA: University of California Press. Quantitative Expression of Cultural Relationships. – 1932. – P. 211–256.
2. Zubin, Joseph. A technique for measuring like-mindedness // The Journal of Abnormal and Social Psychology. – 1938.–P. 508–516.

References

1. Driver and Kroeber (1932). "Quantitative Expression of Cultural Relationships". University of California Publications in American Archaeology and Ethnology. Berkeley, CA: University of California Press. Quantitative Expression of Cultural Relationships: – P. 211–256.
2. Zubin, Joseph (1938). "A technique for measuring like-mindedness". The Journal of Abnormal and Social Psychology. – P. 508–516.

3. Tryon, Robert C. Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality // Edwards Brothers.–1939.
4. Cattell, R. B. The description of personality: Basic traits resolved into clusters // Journal of Abnormal and Social Psychology. – 1943.– P. 476–506.
5. Piatetsky-Shapiro, Gregory. Discovery, analysis, and presentation of strong rules // Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases. – AAAI/MIT Press, Cambridge, MA. – 1991.
6. Agrawal, R. Imieliński, T. Swami. Mining association rules between sets of items in large databases // Proceedings of the 1993 ACM SIGMOD international conference on Management of data – SIGMOD '93. – 1993.– P. 207.
7. Garcia, Enrique. Drawbacks and solutions of applying association rule mining in learning management systems // Sci2s. Archived (PDF) from the original. – 2007.
8. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities // Proc. Natl. Acad. Sci. U.S.A. – 1982. – P. 2554–2558.
9. Neural Net or Neural Network // Gartner IT Glossary. URL: www.gartner.com.
10. Travis Oliphant. Python for Scientific Computing // Computing in Science and Engineering. – 2007.
11. Charles R Harris; K. Jarrod Millman; Stéfan J. van der Walt; et al. Array programming with NumPy (PDF) // 585 (7825): 357–362. ISSN 1476-4687. – 2020.

3. Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.
4. Cattell, R. B. (1943). "The description of personality: Basic traits resolved into clusters". Journal of Abnormal and Social Psychology. – P. 476–506.
5. Piatetsky-Shapiro, Gregory (1991), Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.
6. Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. – P. 207.
7. Garcia, Enrique (2007). "Drawbacks and solutions of applying association rule mining in learning management systems" (PDF). Sci2s. Archived (PDF) from the original.
8. Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational abilities". Proc. Natl. Acad. Sci. U.S.A. – P. 2554–2558.
9. "Neural Net or Neural Network – Gartner IT Glossary". www.gartner.com.
10. Travis Oliphant (2007). "Python for Scientific Computing" (PDF). Computing in Science and Engineering.
11. Charles R Harris; K. Jarrod Millman; Stéfan J. van der Walt; et al. (2020). "Array programming with NumPy" (PDF). // 585 (7825): 357–362.

Надійшла до друку 20.09.2022