УДК 004.02
DOI https://doi.org/10.32782/2663-5941/2023.3.1/37

Фуріхата Д.В.
Державний університет «Житомирська політехніка»

Граф М.С.
Державний університет «Житомирська політехніка»

АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ ТА АЛГОРИТМІВ ОБРОБКИ ІНФОРМАЦІЇ В ІНТЕРНЕТ ПРОСТОРІ

Стаття присвячена дослідженню методів та алгоритмів обробки інформації а Інтернеті. Досліджено технології веб-скрапінгу та API як засоби збору даних з Інтернету. Розглянуто принцип їх роботи, переваги та недоліки та приклади використання. Зосереджено увагу на обробці даних, аналізі, структуризації, кластеризації та класифікації, які є найпоширенішими методами роботи зі зібраними даними. У статті розкрито основні концепції та принципи класичного навчання, яке базується на точно визначених мітках та правильних відповідях, та доведено його ефективність у випадках, де наявні чітко визначені мітки або правильні відповіді. Досліджено різні методи класичного навчання, включаючи лінійну та нелінійну класифікацію, регресію та дерева рішень, а також визначено їх переваги та обмеження. Зазначено, що в мережевому середовищі штучний інтелект, зокрема машинне навчання, виявляється найбільш перспективним для обробки інформації. Подальше дослідження фокусувалося на методах кластеризації даних, які знаходять широке застосування в різних областях. Розглянуто метод k-середніх, який є найпоширенішим та добре вивченим серед методів кластеризації. Визначено, що він мінімізує спотворення та дозволяє ідентифікувати кластери на основі їхніх центроїдів. Однак, було виявлено обмеження методу k-середніх, такі як чутливість до початкової конфігурації центроїдів та можливість збігу до локального мінімуму. Застосування класичного навчання та методів кластеризації даних в машинному навчанні було детально досліджено, а їхню ефективність було підтверджено в різних задачах аналізу та обробки даних. В результаті проведених досліджень встановлено, що використання цих методів може привести до точних та надійних результатів. Таким чином, дана стаття розширює наше розуміння класичного навчання та методів кластеризації даних і надає важливі вказівки для їхнього успішного застосування в практичних задачах машинного навчання.

Ключові слова: *методи, алгоритми, обробка інформації, аналіз даних, машинне навчання.*

Постановка проблеми. Інтернет є необхідною складовою багатьох сфер людської діяльності, включаючи науку, бізнес, культуру, політику та інше. Разом з тим, в даній мережі щодня генерується велика кількість даних, які можуть бути використані для розв'язання різноманітних завдань. Однак, збір та аналіз цих даних може бути складним завданням, тому що вони можуть бути некоректні, неповні, містити шум або бути зібраними в різний час та з різних джерел. Для ефективної обробки цієї інформації необхідні математичні методи та алгоритми аналізу даних. Основні математичні методи обробки інформації в інтернет-просторі включають в себе широкий спектр алгоритмів і технік, які застосовуються для аналізу, фільтрації, класифікації та структурування знань з великого обсягу даних присутніх в Інтернеті. Один з основних методів обробки інформації в Інтернеті – це машинне навчання, що базується на статистичних моделях та алгоритмах, які здатні автоматично вчитись та покращуватись з досвідом. Машинне навчання дозволяє аналізувати дані з Інтернету, виявляти патерни і структури та здійснювати прогнозування та класифікацію.

Аналіз останніх досліджень і публікацій. Згідно з роботою Марченко О.О. та Россади Т.В. [1] існує декілька підходів до побудови моделей даних, а саме: статистичні (базуються на теорії та зосереджуються на перевірці гіпотез), моделі на основі машинного навчання (класичні, нейромережеві, ансамблеві) та обчислювальні (на основі інтелектуального аналізу даних). У дослідженні Олійник А.В. [2] описуються основні етапи збору та аналізу інформації: очищення, інтеграція, вибір параметрів, трансформація, Data Mining (використання інтелектуальних методів для пошуку скритих закономірностей), оцінка та візуалізація. Моделі на основі машинного навчання застосовуються в різних сферах. Авторами роботи [3] запропоновано програмну реалізацію з використанням методів штучного інтелекту для пошуку вільного місця на парковці. Розроблена система надає змогу

користувачам авто здійснювати пошук вільних місць, витрачаючи при цьому мінімум часу. Марчук Д.К. та інші [4] описують процес застосування нейронних мереж для розпізнавання дактильної мови української абетки. В роботі використано особливий різновид архітектури рекурентних нейронних мереж, здатний до навчання, а саме модель довготривалої короткочасної пам'яті LSTM (Long short-term memory), який довів свою ефективність. У статті [5] описано основні алгоритми аналізу потоку кадрів відеоданих, шо надходять з камер міста. Основною метою дослідження є мінімізація часу на пошук вільного місця для паркування автомобіля. У статті [6] досліджуються алгоритми інтелектуального аналізу даних, які на основі правил і обчислень дозволяють створити модель, що аналізує дані, здійснюючи пошук певних закономірностей і тенденцій. Шляхом дослідження алгоритмів інтелектуального аналізу даних було розроблено моделі та методи для встановлення впливу одних хронічних захворювань на інші. Проведені дослідження свідчать про перспективність використання методів інтелектуального аналізу даних для підвищення якості медичної допомоги пацієнтам. В своєму дослідженні Кравченко С.М., Гришкун Є.О. та Власенко О.В. [7], детально описали основні алгоритми класифікації даних та методи підбору математичної моделі обчислень. Основні принципи роботи Байєсівського класифікатору для обробки інформації описані в магістерській дисертації Рудзевич А.П. [8]. В статті Ситника В.Ф.[9] розглянуті практичні приклади реалізації методів дерев рішень (графів) у бізнесі, наукових дослідженнях та фінансах по відношенню до обробки інформації, зокрема і в інтернет-просторі. Описані основні структури, та проведене прогнозування щодо майбутнього розвитку нових алгоритмів навчання. Ткаченко О.М., Біліченко Н.О., Грійо О.Ф., Дзісь О.В. провели дослідження [10] методу k-середніх, та розглянули один із варіантів розв'язку задачі кластеризації даних. Було проведено науковий експеримент та виявлені недоліки та шляхи їх вирішення. Даний метод, при своїй простоті та гнучкості, має низьку швидкість та високу ймовірність сходження до локального мінімуму цільової функції. Авторами була запропонована низка модифікацій: оптимізація при роботі із статичними центроїдами, використання ітеративних алгоритмів та використання kd-дерев.

Метою статті є вивчення принципів та підходів до навчання моделей машинного навчання, а також аналіз важливості збору, підготовки та обробки вхідних даних для досягнення якісних результатів. Дослідження перспектив розвитку методів машинного навчання, зокрема методу k-середніх.

Виклад основного матеріалу. Обробка інформації є ключовим етапом в багатьох дослідницьких проектах, аналітиці, маркетингових дослідженнях та інших сферах, де необхідно отримати об'єктивну та релевантну інформацію. Першим етапом є збір даних. Він має вирішальне значення для отримання якісних результатів та досягнення поставлених цілей. Для цього можуть використовуватися такі технології як веб-скрапінг (англ. web scraping) та API (англ. Application Programming Interface). Веб-скрапінг це технологія, яка полягає в автоматичному зборі даних з веб-сайтів. Він використовується для отримання великих об'ємів даних з різних джерел в інтернеті, зокрема з соціальних мереж, новинних порталів, інтернет-магазинів та інших веб-ресурсів. Для збору даних використовуються спеціальні програми, які називаються скраперами або павуками. Вони працюють за допомогою HTTP-запитів до веб-сторінок, зчитуючи HTML-код і видобуваючи з нього потрібну інформацію. Іноді скрапери використовують технології штучного інтелекту та навчання з підкріпленням, щоб автоматизувати процес вибору необхідної інформації. Іншим методом отримання даних з інтернету є технологія API (англ. Application Programming Interface) – інтерфейс програмування додатків, що дозволяє різним програмним системам взаємодіяти між собою. API надає стандартний спосіб для інтерактивної комунікації між різними програмними системами, дозволяючи їм обмінюватися даними, запитами та відповідями. Дану технологію можна уявити як міст між двома програмними системами. Вона надає засіб для передачі даних між ними, а також правила для того, як ці дані повинні бути передані. API може бути використаний для доступу до різних функцій та сервісів, таких як соціальні мережі, веб-служби, онлайн-магазини, різноманітні сервіси мережевої інфраструктури та ін. API може мати різні форми та протоколи передачі даних. Найбільш поширеними формами API є REST (Representational State Transfer) та SOAP (Simple Object Access Protocol). REST API дозволяє отримувати доступ до ресурсів в мережі за допомогою HTTP-запитів GET, POST, PUT та DELETE. SOAP API передає дані у форматі XML, використовуючи HTTP-протокол. Збір даних є лише першим етапом, після отримання масиву інформації виконується обробка даних, аналіз, структуризація, кластеризація, класифікація та подальше зберігання. Найпоширенішими

методами, що використовуються при обробці інформації в інтернет-просторі є штучний інтелект, а саме машинне навчання. На рис. 1 зображена структура методів штучного інтелекту.

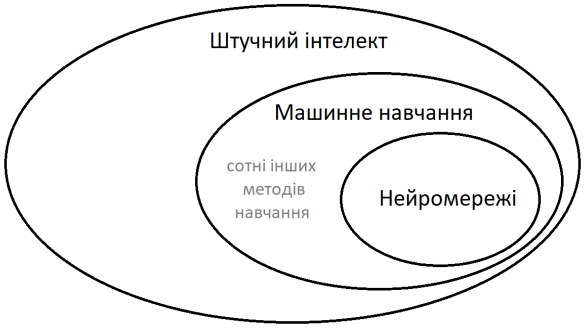

Рис. 1. Структурна схема методів штучного інтелекту

Основна мета машинного навчання полягає у прогнозуванні результатів на основі вхідних даних. Чим більш різноманітні дані ми маємо, тим легше для машини виявити закономірності і отримати точніші результати. Для успішного машинного навчання потрібні три складові: дані, ознаки, алгоритм. Для досягнення кращих результатів необхідна наявність великого обсягу різноманітних даних. Відбір оптимальних ознак є важливим етапом у процесі навчання. Зазвичай цей етап вимагає значних зусиль та займає більше часу, порівняно з іншими етапами навчання моделі. Однак існують ситуації, коли користувач самостійно вирішує, які ознаки вважати «правильними» на основі його власного досвіду та експертної думки. В таких випадках, впровадження суб'єктивних критеріїв може призвести до недостовірних результатів моделювання. У таких ситуаціях модель може видаляти необхідні залежності та навіть втратити здатність адекватно передбачати або аналізувати дані, що викликає появу неправдивих результатів та спотворення інформації. Вибір алгоритму має значний вплив на ефективне вирішення однієї і тієї ж задачі. Правильний вибір методу може впливати на точність, швидкість та обсяг готової моделі. Проте варто пам'ятати, що, незалежно від використовуваного алгоритму, якість вхідних даних має вирішальне значення. Навіть найкращий алгоритм не зможе ефективно працювати з некоректними або непридатними даними. Тому важливо уникати зацикленості на відсотках точності, а зосередитися на зборі якісних даних у максимально можливій кількості.

Існує декілька основних напрямів машинного навчання: класичне навчання, глибинні мережі, навчання з підкріпленням. Класичне навчання є фундаментальним підходом у галузі машинного навчання. Цей підхід базується на використанні наявних даних для тренування моделей з метою отримання точних та універсальних прогнозів або класифікації. В класичному навчанні є всього два напрями, які в свою чергу розгалужуються (рис. 2).

Класичне навчання використовується для побудови моделей на основі точно визначених правильних відповідей або міток. Використання методів класичного навчання має свої сильні

Рис. 2. Напрями класичного навчання

та слабкі сторони, а вибір їх застосування залежить від конкретних ситуацій. Класичне навчання демонструє високу ефективність, коли для набору даних наявні чітко визначені мітки або правильні відповіді. Це дозволяє побудувати точні моделі, які можуть класифікувати нові дані з високою достовірністю. У класичному навчанні використовуються алгоритми, такі як дерева рішень або логістична регресія, які зазвичай дають інтерпретовані результати. Це означає, що людина може аналізувати та розуміти причинно-наслідкові зв'язки, які лежать в основі прийнятих моделей. Класичне навчання потребує чітко визначених міток або правильних відповідей для навчальних даних. Якщо мітки недостовірні, неточні або недоступні, це може призвести до поганих результатів моделі і неправильних висновків. Отже, точність і надійність мають значну залежність від якості міток. Методи класичного навчання рідко використовуються при роботі з великими обсягами даних. Обробка та аналіз великих наборів даних може стати витратною за ресурсами та часом, особливо при використанні складних алгоритмів. Це може вплинути на продуктивність моделі та затримати процес навчання. Класичне навчання ефективне в ситуаціях, де доступні якісні та достовірні мітки або правильні відповіді, а також коли модель може бути побудована на основі цих даних.

Одним із завдань машинного навчання це кластеризація. Кластеризація даних є відомою задачею як у наукових, так і в практичних сферах. Її основна мета полягає в розподілі експериментально отриманих наборів векторів на групи, відомі як кластери. Кластеризація широко використовується в статистичному аналізі даних, векторній квантизації, розпізнаванні образів та інших областях. В галузі ущільнення мовлення алгоритми кластеризації використовуються для створення кодових книг, які містять найбільш репрезентативні набори даних. Задачу кластеризації можна сформулювати наступним чином: виходячи з заданого набору з n векторів розмірності d, необхідно розбити їх на підмножини згідно з певним критерієм оптимізації. Зазвичай, таким критерієм є мінімізація спотворення. Метод кластеризації k-середніх є найбільш поширеним і добре вивченим серед усіх методів кластеризації. Він мінімізує спотворення, розподіляючи дані між непересетинаючими регіонами і ідентифікуючи їх за їхніми центрами. Головними перевагами методу k-середніх є його простота, гнучкість та швидка збіжність. Однак цей метод має обмеження, такі як значна залежність результатів кластеризації від початкової конфігурації центроїдів (ініціалізації), повільна обробка великих обсягів даних та можливість збігу до локального мінімуму цільової функції. Існують різні модифікації методу k-середніх. Наприклад, метод k-середніх++ використовує вдосконалену процедуру ініціалізації, що дозволяє отримати кращі результати кластеризації шляхом спеціального вибору початкової конфігурації центроїдів. Інші підходи включають відкидання статичних центроїдів для прискорення обчислення відстаней і застосування ітеративного алгоритму для наближення до глобального оптимуму шляхом послідовного запуску k-середніх.

Метод k-середніх полягає в тому, що спочатку вибирається деяка кількість кластерів (k), після чого об'єкти розподіляються між цими кластерами залежно від їх відстані до центроїдів (середніх значень) кожного кластера. Потім центроїди перераховуються, і процес розподілу об'єктів повторюється до тих пір, поки кластери стабілізуються. У кластеризації даних важливо враховувати такі фактори, як відстань між об'єктами, кількість кластерів та критерії оцінки якості кластеризації.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left(x_j - \mu_i \right)^2 \qquad (1),$$

де k – число кластерів, S_i – отримані кластери, i = 1 , 2 , … , k, μ_i – центри мас векторів x_j∈S_i.

Кластеризація даних також може використовуватися для вирішення різноманітних задач, таких як аналіз відгуків клієнтів, виявлення аномальних даних, класифікація образів, рекомендації товарів та багато інших.

Висновки. Машинне навчання є поширеним методом обробки інформації в Інтернеті. Воно використовує дані, ознаки і алгоритми для прогнозування результатів.

Класичне навчання – основний напрямок машинного навчання, який базується на використанні наявних даних для побудови точних моделей. Даний напрям має сильні сторони, такі як висока ефективність при наявності чітко визначених міток або правильних відповідей для навчальних даних. Він використовує інтерпретовані алгоритми, що дозволяють аналізувати та розуміти причинно-наслідкові зв'язки. Кластеризація даних полягає в розподілі об'єктів на групи, відомі як кластери, і має різноманітні застосування. Кластеризація має широкий спектр застосувань, включаючи аналіз відгуків клієнтів, виявлення аномалій, класифікацію образів, рекомендації товарів та багато інших. Цей підхід допомагає вирішувати різноманітні задачі і виявляється корисним інструментом у процесі обробки даних. Загалом,

стаття надає важливі відомості про методи штучного інтелекту, зокрема машинного навчання, і їх застосування в обробці інформації. Крім того, досліджується задача кластеризації даних та використання методу k-середніх у цьому контексті. Отримані висновки можуть бути корисними для дослідників і практиків, які займаються областями штучного інтелекту і аналізу даних.

Список літератури:

1. Марченко О.О., Россада Т.В. «Актуальні проблеми Data Mining: Навчально-методичний посібник». Київський національний університет імені Тараса Шевченка, 2017.
2. Олійник А.В. Data Mining як інструмент оцінки та прогнозування найвагоміших показників ефективності сторінок брендів в соціальних мережах// Магістерська робота. – Київ: Національний Університет Києво-Могилянська академія, 2020. URL: https://ekmair.ukma.edu.ua/server/api/core/bitstreams/ba2064e0-844e-4785-b209-980deb51c495/content
3. Левківський В.Л., Марчук Г.В., Ципоренко В.В., Марчук Д.К. Комп'ютерна програма «Алгоритмічно-програмне забезпечення обробки та аналізу потоку кадрів відеоданих, що надходять з камер міста». – 2021. URL: http://eztuir.ztu.edu.ua/bitstream/handle/123456789/8019/109822.pdf?sequence=1&isAllowed=y
4. Марчук Д.К., Левківський В.Л., Марчук Г.В., Голенко М.Ю. Система розпізнавання дактильної мови української абетки. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки. – 2022. – Том 33 (72), № 6. – С. 109-114. URL: https://doi.org/10.32782/2663-5941/2022.6/19
5. Levkivskyi V., Marchuk, Lobanchykova N., Pilkevych I., Salamatov D. Available parking places recognition system. CEUR Workshop Proceedings 4th Workshop for Young Scientists in Computer Science & Software Engineering. Volume 3077 (2022). pp. 123-134. URL: http://ceur-ws.org/Vol-3077/paper07.pdf
6. Левківський В., Лобанчикова Н., Марчук Д. Дослідження алгоритмів Data Mining // E3S Web of Conferences. 2020. Том 166. С. 05007. https://doi.org/10.1051/e3sconf/202016605007.
7. Кравченко С.М., Гришкун Є.О., Власенко О.В. «Методи класифікації машинного навчання з використанням бібліотеки scikit-learn». *Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки.* 2020. URL: http://tech.vernadskyjournals.in.ua/journals/2020/3_2020/part_1/21.pdf
8. Рудзевич І.О. Методи машинного навчання в сентимент аналізі текстової інформації // Магістерська робота. – Київ: Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», 2020. URL: https://ela.kpi.ua/bitstream/123456789/35699/1/Rudzevich_magistr.pdf
9. Ситник В.Ф. , Ситник Н.В,. Дерева рішень в системах дейтамайнінгу, Survey on Data Mining Techniques for Smart Grids // International Journal of Advanced Computer Science and Applications. – 2019. – Vol. 10, No. 7. – P. 106-112. URL: https://core.ac.uk/download/pdf/197259458.pdf
10. Ткаченко О.М., Біліченко Н.О. Метод кластеризації на основі послідовного запуску к-середніх з обчисленням відстаней до активних центроїдів. // Вінницький національний технічний університет. URL: https://dspace.nbuv.gov.ua/bitstream/handle/123456789/50557/03-Tkachenko.pdf?sequence=1

Fyrikhata D.V., Graf M.S. ANALYSIS OF EXISTING METHODS AND ALGORITHMS FOR INFORMATION PROCESSING IN THE INTERNET SPACE

The article is dedicated to investigating methods and algorithms for information processing on the Internet. The technologies of web scraping and APIs as means of data collection from the Internet have been examined. The principles of their operation, advantages, disadvantages, and examples of their usage have been considered. Attention has been focused on data processing, analysis, structuring, clustering, and classification, which are the most common methods for working with collected data. The article reveals the fundamental concepts and principles of classical learning, which are based on precisely defined labels and correct answers, and demonstrates its effectiveness in cases where clear labels or correct answers are available. Various methods of classical learning have been explored, including linear and nonlinear classification, regression, and decision trees, and their advantages and limitations have been determined. It has been noted that in the networked environment, artificial intelligence, particularly machine learning, proves to be the most promising for information processing. Further research has been centered on data clustering methods. The k-means method, being the most common and well-studied clustering method, has been examined. However, limitations of the k-means method have been identified, such as sensitivity to the initial configuration of centroids and the possibility of convergence to a local minimum. The application of classical learning and data clustering methods in machine learning has been thoroughly investigated, and their effectiveness has been confirmed in various data analysis and processing tasks. As a result of the conducted research, it has been established that the utilization of these methods can lead to accurate and reliable results. Thus, this article expands our understanding of classical learning and data clustering methods and provides important guidelines for their successful application in practical machine learning tasks.

Key words: *methods, algorithms, information processing, data analysis, machine learning.*