УДК 004.422.5

Засипка Є. В., здобувач 2 курсу
спеціальності 122 Комп'ютерні науки,
Павлюк О. А., канд. техн. наук, старший
викладач кафедри інформаційних технологій

АЛГОРИТМ ОБРОБКИ ДАНИХ У ШТУЧНОМУ ІНТЕЛЕКТІ

Донецький національний університет імені Василя Стуса, м. Вінниця

Алгоритм обробки даних у штучному інтелекті – це чітко визначена послідовність інструкцій та процедур, спрямованих на перетворення сирих, неструктурованих даних у чистий, структурований та анотований формат, придатний для ефективного навчання моделей штучного інтелекту [1]. Цей алгоритм є критично важливим конвеєром попередньої підготовки, який забезпечує якість вхідної інформації, а отже, і кінцеву продуктивність і надійність інтелектуальної системи [2].

Алгоритм обробки даних у штучному інтелекті є фундаментальним процесом, що перетворює сирі неструктуровані дані на чисту та придатну для навчання моделей інформацію. Актуальність цієї теми зумовлена тим, що якість будь-якої інтелектуальної системи безпосередньо залежить від якості вхідних даних, відтак алгоритм їх обробки виступає критично важливою ланкою у конвеєрі розробки штучного інтелекту [1].

Основним завданням алгоритму обробки даних є підготовка інформаційного масиву, який дасть змогу моделі штучного інтелекту ефективно виявляти закономірності та ухвалювати точні прогнози. Цей процес можна розглядати як послідовність взаємопов'язаних етапів. Першим кроком завжди є збір та отримання даних із різноманітних джерел, серед яких можуть бути структуровані бази даних, текстові документи, зображення, аудіозаписи та потоки інформації з сенсорів. Ключовою вимогою на цьому етапі є забезпечення релевантності та достатнього обсягу вибірки [3].

Наступна фаза, попередня обробка та очищення, є найбільш трудомісткою. Вона включає в себе низку важливих процедур, спрямованих на підвищення якості даних. Необхідно вирішити проблему пропущених значень шляхом їх видалення, наприклад, заповненням середніми значеннями. Кодування категоріальних змінних перетворює текстовий опис на числовий формат за допомогою методів, як от One-Hot Encoding. Критично важливим є видалення дублікатів та ідентифікація аномальних значень (викидів), які можуть суттєво спотворити результат навчання. Також дані часто потребують нормалізації або стандартизації для приведення всіх числових ознак до єдиного масштабу, що значно пришвидшує процес навчання моделі [2].

Фінальним етапом підготовки часто виступає відбір ознак, оскільки не всі змінні однаково корисні для моделі. Надмірна кількість ознак веде до перенавчання, коли модель запам'ятовує шум замість реальних закономірностей. Методи відбору варіюються від простих фільтраційних, що оцінюють статистичну зв'язність ознаки з цільовою змінною, до складніших вкладених методів, які оціню-

ють важливість ознаки на основі продуктивності самої моделі. Варто зазначити, що специфіка алгоритму обробки значно відрізняється для різних типів даних. Для зображень це зведення до одного масштабу яскравості пікселів та аугментація; для природної мови це токенізація, видалення стоп-слів і лематизація; для часових рядів – видалення тренду та сезонності [2].

Окремої уваги заслуговують проблеми, пов'язані з якістю самих вихідних даних. Алгоритми обробки не можуть усунути системні упередження, які вже містяться в даних. Наприклад, якщо історичні дані про прийом на роботу відображають певну соціальну нерівність, модель лише відтворить і посилить це упередження. До того ж складність масштабування становить серйозну проблему – методи, що ефективно працюють на невеликих наборах даних, часто виявляються непридатними для роботи з великими даними через обмеження обчислювальних ресурсів. Отже, навіть ідеально виконаний алгоритм обробки даних не може гарантувати успіх, якщо вихідні дані містять фундаментальні вади. Це підкреслює той факт, що обробка даних залишається не лише технічною, але значною мірою методологічною та навіть філософською проблемою, де інтуїція та розуміння предметної області часто виявляються важливішими за формальні математичні процедури [3].

У висновку можна констатувати, що алгоритм обробки даних є не допоміжним, а центральним процесом у створенні ефективного штучного інтелекту. Це ітеративний і часто творчий ланцюг дій, що перетворює хаотичну інформацію на структуроване знання. Майбутній розвиток цієї галузі, ймовірно, буде пов'язаний з подальшою автоматизацією всього конвеєра обробки в межах концепції AutoML, а також із розробкою ефективніших методів роботи з великими та невпорядкованими даними. Інвестиції в якісну підготовку даних завжди є найбільш ефективною стратегією для побудови точних, надійних та робочих систем штучного інтелекту [2].

Список використаних джерел

1. ДСТУ 2860:2014. Системи штучного інтелекту. Терміни та визначення. На заміну ДСТУ 2860-94. Київ, 2014. URL: https://online.budstandart.com/ua/catalog/doc-page.html?id_doc=25034 (дата звернення: 22.11.2025).

2. Russell S. J., Norvig P. Artificial intelligence: a modern approach. 4th ed. Hoboken: Pearson. 2020.

3. Федорін І. В. Проектування та аналіз обчислювальних алгоритмів: вступ до алгоритмів: навчальний посібник. Київ: КПІ ім. Ігоря Сікорського. 2022.