

ВИКОРИСТАННЯ LLM ДЛЯ АВТОМАТИЗОВАНОГО АНАЛІЗУ ЯКОСТІ ТАБЛИЧНИХ ДАНИХ

Кірюшин В. А.

e-mail: volodymyr.kiriushyn@nure.ua

Науковий керівник – к.т.н, доц. Вітько О.В.

Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This paper introduces AEDA (Automated Exploratory Data Analysis), a framework for automated data-quality assessment in tabular machine-learning datasets. AEDA integrates diverse analytical modules, including statistical analysis, missing-value and outlier detection, and rule mining, to evaluate data integrity. Uniquely, it employs a lightweight local large language model to generate explanations for identified issues and computes an overall quality score in accordance with ISO/IEC 25012 [1]. Experimental evaluations on real-world datasets confirm the framework's scalability and practical effectiveness.

У сучасних задачах машинного навчання та аналізу даних якість вихідних даних відіграє ключову роль. Навіть найбільш складні алгоритми не здатні забезпечити коректні результати у випадку використання неповних, зашумлених або неконсистентних наборів даних. Саме тому етап попереднього аналізу та перевірки якості даних є важливою частиною процесу їх підготовки.

Традиційно дослідження структури даних здійснюється за допомогою так званого дослідницького аналізу даних (Exploratory Data Analysis, EDA) [2]. Однак, виконання такого аналізу вручну може вимагати значних часових витрат, особливо при роботі з великими табличними наборами даних. У зв'язку з цим актуальною задачею є створення інструментів, які дозволяють автоматизувати процес аналізу якості даних.

Одним з таких рішень є програмний фреймворк AEDA (Automated Exploratory Data Analysis), реалізований мовою Python. Основною метою даного підходу є автоматичне виконання комплексної перевірки табличних наборів даних з використанням набору спеціалізованих модулів аналізу.

У рамках фреймворку реалізовано одинадцять компонентів аналізу даних. До них відносяться базові статистичні перевірки, аналіз пропущених значень, пошук дублікатів, виявлення аномальних значень, а також визначення потенційно помилкових значень цільової функції у задачах класифікації. Додатково застосовуються методи виявлення закономірностей між атрибутами та алгоритми пошуку аномалій на основі автоенкодерів, що широко використовуються у задачах машинного навчання [3].

Окремою особливістю системи є використання великої мовної моделі (Large Language Model, LLM) для формування текстових пояснень

результатів аналізу. Для цього використовується компактна модель, яка може працювати локально без необхідності підключення до хмарних сервісів. Модель аналізує результати перевірок та формує короткі інтерпретації виявлених проблем, що значно полегшує роботу аналітиків та інженерів даних.

Для комплексної оцінки якості набору даних у фреймворку реалізовано систему підрахунку інтегрального показника якості. В основі цього показника лежать характеристики стандарту ISO/IEC 25012, який описує основні властивості якості даних, зокрема, повноту, узгодженість, точність та актуальність [4].

Експериментальна перевірка запропонованого підходу проводилась на декількох реальних наборах даних. Зокрема, було використано відомий навчальний набір Titanic, а також великий даних ресторанів міста Бангалор. Результати експериментів показали, що запропонований фреймворк ефективно виявляє різні типи проблем якості даних та може масштабуватися для роботи з великими обсягами інформації.

Таким чином, використання автоматизованих інструментів аналізу якості даних дозволяє суттєво скоротити час підготовки даних для задач машинного навчання. Як показують сучасні дослідження, перехід від суто орієнтованого на дані до знання-орієнтованого штучного інтелекту за допомогою LLM значно підвищує адаптивність систем [5]. Тому інтеграція методів аналізу даних з можливостями великих мовних моделей відкриває нові перспективи для створення інтелектуальних систем підтримки роботи з даними

Список використаних джерел:

1. ISO/IEC 25012:2008. Data quality model. Geneva: International Organization for Standardization, 2008.
2. Tukey J. W. Exploratory data analysis. Reading : Addison-Wesley, 1977. 688 p.
3. Goodfellow I., Bengio Y., Courville A. Deep learning. Cambridge : MIT Press, 2016. 775 p.
4. Han J., Kamber M., Pei J. Data mining: concepts and techniques. 3rd ed. Waltham: Morgan Kaufmann, 2012. 703 p.
5. Terziyan V., Vitko O., Terziyan O., Terziian A. Hybrid Intelligence for Industry 4.0: Integrating Large Language Models and Reinforcement Learning. International Journal of Computer Information Systems and Industrial Management Applications. 2025. Vol. 17. P. 610–627. DOI: <https://doi.org/10.70917/ijcisim-2025-0038>.