НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ПОЛТАВСЬКА ПОЛІТЕХНІКА
ІМЕНІ ЮРІЯ КОНДРАТЮКА

ЗБІРНИК МАТЕРІАЛІВ

77-Ї НАУКОВОЇ КОНФЕРЕНЦІЇ ПРОФЕСОРІВ, ВИКЛАДАЧІВ, НАУКОВИХ ПРАЦІВНИКІВ, АСПІРАНТІВ ТА СТУДЕНТІВ УНІВЕРСИТЕТУ

16 травня – 22 травня 2025 р.

УДК 004.058

О.С. Скорбатюк, асистент
М.О. Толочин, асистент
А.О. Мізік, асистент
Національний університет
«Полтавська політехніка імені Юрія Кондратюка»

МЕТОДИ ПОЯСНЮВАНОСТІ ANFIS-ПОДІБНИХ МЕРЕЖ

В умовах стрімкого розвитку штучного інтелекту дедалі важливішою стає потреба в забезпеченні прозорості та зрозумілості його рішень. Пояснювальний штучний інтелект (XAI, Explainable Artificial Intelligence) відіграє ключову роль у підвищенні довіри до автоматизованих систем, особливо в критично важливих галузях, таких як медицина, фінанси, оборона [1]. Здатність пояснити, чому була прийнята певна модель поведінки або прогноз, є необхідною умовою для інтеграції ШІ у реальні процеси прийняття рішень, забезпечуючи відповідальність, надійність та безпеку.

Адаптивна нейро-нечітка система виведення (ANFIS) поєднує навчальні можливості нейронних мереж з інтерпретованістю нечітких правил, що робить її перспективною в контексті пояснювального ШІ. Завдяки зрозумілій структурі та здатності моделювати складні залежності, доцільно проаналізувати підходи до пояснення роботи таких мереж.

Можна виділити два підходи до пояснюваності: ante-hoc та post-hoc. Ante-hoc полягає в закладенні інтерпретованості та поясності в саму структуру моделі, тоді як post-hoc ставить ціллю досягнення поясності вже після того, як модель була створена без втручання та внесення змін в її структуру.

ANFIS є ante-hoc моделлю, оскільки в її основі лежать нечіткі правила, які є природно інтерпретованими людиною, а її проста та прозора структура чітко визначає ролі для кожного шару.

Варто відзначити те, що ante-hoc дизайн не гарантує пояснюваності результатів [2]. Так, наприклад, хоча структура ANFIS і є прозорою, зі збільшенням кількості входів інтерпретованість знижується через ріст правил і параметрів. Іншою проблемою є те, що під час навчання мережі відбувається підгонка функцій належності, коефіцієнтів правил, що може призвести до втрати ними їх оригінального значення.

У зв’язку з обмеженнями ante-hoc підходів щодо масштабованості та збереження інтерпретованості, постає доцільним використання post-hoc методів, для компенсації втрати інтерпретованості під час навчання.

Модель-агностичні post-hoc методи аналізують лише вхідні та вихідні дані, що дозволяє застосовувати їх до широкого кола моделей, зокрема і до ANFIS. Прикладами таких методів є статистичні методи, методи вилучення правил, моделей – LIME, SHAP тощо [3].

Модель-специфічні методи, навпаки, використовують внутрішню інформацію моделі для пояснення прийнятих рішень. Прозорість ANFIS є перевагою, оскільки чітка організація шарів полегшує інтерпретацію. Для глобального пояснення достатньо аналізу функцій належності та нечітких правил, а для локального аналізу самих кроків шляху обчислення: активації функцій належності, силу спрацювання правил та їх результатів.

Окремим важливим аспектом як XAI загалом, так і стосовно ANFIS-подібних мереж зокрема, є формування і представлення пояснень у такому вигляді, який буде зрозумілим користувача, що є ключовою вимогою для повноцінної оцінки прийнятого системою рішення.

Одним із перспективних підходів до вирішення цієї проблеми є використання LLM для трансформації структурованих результатів інтерпретації у природну мову. Мовна модель може як просто підводити підсумки, так і бути повноцінною експертною системою, надаючи довідкову інформацію, аналітичні оцінки, рекомендації та інші форми пояснень [4].

Підсумовуючи, ANFIS-подібні мережі, завдяки поєднанню прозорості нечітких систем і адаптивності нейронних мереж, мають значний XAI потенціал. Для забезпечення пояснюваності є необхідним використання як ante-hoc, так і post-hoc методів. Особливо перспективним є поєднання структурованої інтерпретації з представленням у вигляді природної мови за допомогою великих мовних моделей, що дозволяє підвищити зрозумілість і прийнятність рішень для кінцевого користувача.

Література:

1. A Literature Review on Applications of Explainable Artificial Intelligence (XAI) / K. Kalasampath et al. IEEE Access. 2025. P. 1. URL: https://doi.org/10.1109/access.2025.3546681 (date of access: 06.05.2025).

2. Sokol K., Vogt J.E. (Un)reasonable allure of ante-hoc interpretability for high-stakes domains: Transparency is necessary but insufficient for comprehensibility [Електронний ресурс] // arXiv [cs.LG]. – 2023. – Режим доступу: https://doi.org/10.48550/ARXIV.2306.02312.

3. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence / V. Hassija et al. Cognitive Computation. 2023. URL: https://doi.org/10.1007/s12559-023-10179-8 (date of access: 06.05.2025).

4. Murillo E. de C. S. Unveiling the black box: The significance of XAI in making LLMs transparent [Електронний ресурс] / E. de C. S. Murillo // Zenodo. – 2025. – Режим доступу: https://doi.org/10.5281/ZENODO.14885027.