УДК 004.7

Бондарєв Андрій Сергійович, аспірант,
П`ятаченко Владислав Юрійович, асистент
Сумський державний університет

АНАЛІЗ МЕТОДІВ ОБРОБКИ НЕПОВНИХ ДАНИХ ДЛЯ ВЕБОРІЄНТОВАНОЇ СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

Актуальність проблеми. Системи підтримки прийняття рішень (СППР) на основі машинного навчання активно використовуються у різних сферах діяння, зокрема в освіті. Однією з критичних проблем при цьому є неповнота даних. Для веборієнтованої СППР для оцінювання освітнього контенту за оцінками експертів проблема проявляється вигляді порожніх полів. Подібна неповнота призводить до зміщення розподілу навчальних даних, погіршення точності передбачень і некоректних висновків.

1. У статті «The impact of imputation quality on machine learning classifiers for datasets with missing values» досліджено взаємозалежність між якістю відновлення пропущених значень і точністю класифікаторів, а також порівняно методи для відновлення даних. Встановлено, що найбільший вплив на якість класифікації має частка пропущених значень у тестовій вибірці, а не лише у навчальній. Автори пропонують новий клас метрик якості імпутації на основі відстані Васерштейна, який значно краще відображає збереження реального розподілу даних порівняно із традиційними метриками. Результати демонструють, що некоректна імпутація погіршує точність та інтерпритованість моделей.

2. У статті від Yoon, Jordon, van der Schaar (2018) розглянуто чи можна оцінити пропущені значення на основі спостережуваних даних за допомогою нейромережевого підходу, що навчається відтворювати справжній розподіл даних.

Автори запропонували метод відновлення пропущених значень на основі генеративно-змагальної мережі. Генератор відновлює пропуски з урахуванням наявних значень, тоді як дискримінатор визначає для кожної ознаки – чи є значення реальним чи відновленим. Таким чином, GAIN демонструє принципову перевагу GAN-підходу над статистичними методами.

3. Автори у роботі «Generative Adversarial Networks for Imputing Sparse Learning Performance» (ICPR) вирішують задачу наступну задачу: є матриця оцінок, де рядки - суб'єкти, стовпці - дисципліни, а більша частина клітинок порожня, бо не кожен студент виконував кожне завдання. Автори адаптують GAN-підхід для таких матричних даних,

додаючи CNN-шари для вловлювання залежностей між сусідніми дисциплінами. Метод валідовано на шести реальних навчальних датасетах і перевершує методи матричної факторизації та оригінальний GAIN. Це безпосередньо обґрунтовує застосування GAN-підходу для відновлення неповних експертних оцінок у веборієнтованій СППР, де природа розрідженості це відсутність предметної компетентності окремих.

Проектне рішення для системи. На основі проведеного аналізу обґрунтовано архітектурне рішення для веборієнтованої СППР. Запропонований pipeline обробки неповних даних включає: (1) визначення механізму пропуску та аналізу патернів відсутності – з урахуванням специфіки предметної галузі, де пропуск зумовлений відсутністю компетентності експерта оцінити дисципліну, а не випадковим фактором; (2) відновлення пропущених значень методом GAIN з CNN-розширенням для матричних даних або MICE як статистичною альтернативою при малій частці пропусків; (3) оцінку якості відновлення за метрикою на основі відстані Васерштейна, що забезпечує контроль збереження розподілу даних;

Висновки. У розглянутій задачі відсутність оцінки зумовлена браком компетентності експерта. Застосування GAN-підходу, адаптованого для матричних освітніх даних, у поєднанні з метричним контролем якості відновлення є перспективним напрямом для побудови інтелектуальної системи підтримки прийняття рішень з неповною визначеністю даних.

Список використаних джерел:

1. The impact of imputation quality on machine learning classifiers for datasets with missing values / T. Shadbahr, M. Roberts, J. Stanczuk et al. Communications Medicine. 2023. Vol. 3, art. 139. DOI: https://doi.org/10.1038/s43856-023-00356-z (дата звернення: 24.03.2026).

2. Yoon J., Jordon J., van der Schaar M. GAIN: Missing Data Imputation using Generative Adversarial Nets. Proceedings of the 35th International Conference on Machine Learning (ICML 2018). 2018. Vol. 80. P. 5689–5698. URL: https://arxiv.org/abs/1806.02920 (дата звернення: 24.03.2026).

3.Generative Adversarial Networks for Imputing Sparse Learning Performance / L. Zhang, H.-Y. Shum, V. Shute, F. Wang. Proceedings of the 27th International Conference on Pattern Recognition (ICPR 2024). 2024. URL: https://arxiv.org/abs/2407.18875 (дата звернення: 24.03.2026).