

РЕЦЕНЗІЯ

на дисертаційну роботу **Земляного Олексія Дмитровича**
на тему **«Розроблення методів та програмного забезпечення
інтелектуального імпутування пропусків даних»**,
представлену на здобуття ступеня доктора філософії
за спеціальністю 121 Інженерія програмного забезпечення

Актуальність теми дисертації. Однією з поширених проблем сучасного аналізу даних є наявність пропусків у наборах даних, що виникають під час збору, передавання та зберігання інформації. Пропущені значення можуть знижувати точність моделей класифікації й прогнозування, обмежувати ефективність використання методів машинного навчання, а також негативно впливати на достовірність результатів аналізу та обґрунтованість рішень, що приймаються на їх основі.

Водночас існуючі методи імпутування не є універсальними, а їх ефективність значною мірою залежить від особливостей даних, типів ознак, механізмів виникнення та патернів пропусків. Це зумовлює необхідність розроблення інтелектуальних методів імпутування, здатних використовувати статистичні закономірності, взаємозв'язки між ознаками та алгоритми машинного навчання для більш точного відновлення пропущених значень.

Практична значущість розроблення таких методів підтверджується широкими можливостями їх застосування у задачах аналізу даних різної природи, зокрема під час обробки медичних даних та даних гідрологічного моніторингу, які використано у дисертаційному дослідженні.

Не менш важливим є створення програмного забезпечення, яке забезпечує ефективну реалізацію таких методів та можливість їх використання у сучасних інформаційно-аналітичних системах. Реалізація методів у вигляді програмних компонентів, сумісних із поширеними бібліотеками машинного навчання, сприяє їх широкому застосуванню, відтворюваності експериментальних досліджень та інтеграції у сучасні конвеєри обробки даних.

Таким чином, розроблення методів та програмного забезпечення інтелектуального імпутування пропусків у даних є актуальною науково-практичною задачею, що має важливе теоретичне і практичне значення для

розвитку інженерії програмного забезпечення в частині створення програмних засобів інтелектуального аналізу даних.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційне дослідження виконувалося в межах науково-дослідних тем «Розроблення програмного забезпечення аналізу та кластеризації часових рядів» (номер державної реєстрації 0122U001465, 2022–2024 рр.) та «Розроблення інформаційної технології обробки статистичних даних» (номер державної реєстрації 0125U002280, 2025–27 рр.). Крім того, дослідження отримало часткове фінансування від Міністерства закордонних справ Чеської Республіки в межах проекту № 24-PKVV-UM-011 «Посилення стандартів викладання, досліджень та міжнародного співробітництва в Дніпровському національному університеті імені Олеся Гончара (ДНУ)», реалізованого Карловим університетом спільно з ДНУ. Це свідчить про тісний зв'язок дисертаційної роботи з науковими програмами і міжнародними науково-освітніми проектами.

Наукова новизна отриманих автором результатів. У дисертаційній роботі розв'язано актуальну наукову задачу, що полягає у розробленні методів, алгоритмів та програмних засобів інтелектуального заповнення пропусків у даних, які забезпечують точність та ефективність у задачах моніторингу. До основних результатів, що визначають наукову новизну роботи, належать:

1. Удосконалено метод імпутування NoNa шляхом розроблення методу UnifiedClassRegrImputer, який використовує комбінацію класифікатора та регресора з урахуванням типу даних ознак, патернів пропусків і змінного порядку їх обробки, що дозволяє підвищити точність імпутування та покращити якість моделей у задачах класифікації та прогнозування.

2. Вперше розроблено метод імпутування RegrImputer, який використовує регресор з уточненням значень для категоріальних даних та дозволяє підвищити швидкодію та точність імпутування, а також покращити якість моделей в задачах класифікації та прогнозування.

3. Удосконалено ентропійний метод EntropyImputer шляхом додавання однокрокової та ітераційної процедур зменшення умовної ентропії кожної ознаки, реалізованих у послідовному та паралельному режимах, що дозволяє зменшити умовну ентропію та час виконання алгоритму.

4. Вперше розроблено гібридний метод імпутування HybridRegrEntropyImputer, який поєднує ентропійний та регресійний підходи, що дозволяє підвищити точність імпутування та швидкість порівняно з EntropyImputer.

5. Вперше розроблено метод імпутування CorrelationImputer, який використовує виявлений кореляційний зв'язок між ознаками та автоматизує вибір найкращої моделі залежності серед лінійних і квазілінійних моделей, що дозволяє підвищити точність імпутування.

6. Вперше запропоновано метод імпутування пропусків у часових рядах з сезонною складовою на основі виявленого кореляційного зв'язку між ознаками (метод CorrelationImputer), у якому імпутування виконується окремо для кожного місяця року, що забезпечує підвищення точності імпутування і якості моделей прогнозування порівняно з аналогом без розбиття за місяцями.

7. Удосконалено методи перетворення якісних ознак на кількісні шляхом збереження інформації про пропуски та забезпечення можливості виконання зворотного перетворення:

- IgnoreNANLabelEncoder – модифікація стандартного LabelEncoder, яка не кодує пропуски, зберігає словник відповідностей і підтримує зворотне перетворення;

- IgnoreNANFrequentEncoder – модифікація LabelEncoder та FrequencyEncoder, яка враховує частоту появи значень, надаючи більшу вагу найпоширенішим або найрідшим категоріям, не кодує пропуски, зберігає словник та підтримує зворотне перетворення.

Практичне значення одержаних результатів полягає у розробленні алгоритмів і програмного забезпечення для імпутування пропусків даних, реалізованих мовою Python відповідно до архітектурних принципів бібліотеки scikit-learn, що забезпечує їх інтеграцію до сучасних конвеєрів обробки даних. Запропоновані методи дозволяють підвищити точність імпутування та покращити якість моделей класифікації і прогнозування, насамперед для даних медичного та гідрологічного моніторингу. Практична цінність результатів підтверджується їх впровадженням в освітній процес кафедри інженерії програмного забезпечення та інформаційних технологій Дніпровського національного університету імені Олеся Гончара, а також можливістю

використання розробленого програмного забезпечення у прикладних системах аналізу даних.

Обґрунтованість і достовірність наукових положень, висновків і рекомендацій, які захищаються, забезпечуються коректною постановкою завдань дослідження, використанням сучасних методів аналізу даних, машинного навчання, математичної статистики та технологій інженерії програмного забезпечення, а також проведенням комплексних порівняльних експериментальних досліджень, які підтверджують ефективність запропонованих методів за показниками точності імпутування, швидкодії та впливу на якість моделей класифікації і прогнозування.

Особистий внесок здобувача. Дисертаційна робота є самостійною науковою працею, у якій викладено власні наукові ідеї та результати розробок автора. Основні положення, висновки та результати, що виносяться на захист, отримано здобувачем особисто або за його безпосередньої участі. Автором самостійно проведено аналіз літературних джерел, розроблено методи та алгоритми для імпутування пропусків у даних, виконано програмну реалізацію запропонованих підходів мовою Python та проведено комплексні порівняльні експериментальні дослідження.

Апробація результатів дисертації. Основні результати дисертаційного дослідження апробовано на міжнародних наукових конференціях «Математичне та програмне забезпечення інтелектуальних систем (MPZIS)» (м. Дніпро, 2023, 2024, 2025), «Автоматика-2024» (м. Дніпро, 2024), «Виклики та проблеми сучасної науки» (м. Дніпро, 2024), «Тенденції та перспективи розвитку науки і освіти в умовах глобалізації» (м. Переяслав, 2023), а також підсумкових наукових конференціях Дніпровського національного університету імені Олеся Гончара (м. Дніпро, 2023, 2024, 2025, 2026). Основні положення дисертації опубліковано у 11 наукових працях, серед яких 4 статі у наукових фахових виданнях України категорії Б, 2 статті у інших збірниках наукових праць та 5 тез доповідей у збірниках матеріалів наукових конференцій.

Структура та обсяг дисертації. Дисертаційна робота складається з анотацій українською та англійською мовами, переліку умовних позначень, вступу, чотирьох розділів, висновків, списку використаних джерел із 121

найменування та трьох додатків. Загальний обсяг дисертації – 169 сторінок, обсяг основного тексту – 117 сторінок.

У вступі дисертаційної роботи розкрито актуальність теми, визначено мету, завдання, об'єкт та предмет дослідження, сформульовано наукову новизну та практичне значення одержаних результатів, наведено відомості про особистий внесок здобувача, апробацію результатів та публікації.

Перший розділ присвячено аналізу сучасних методів імпутування пропусків у даних. У розділі наведено класифікацію пропусків у даних, проаналізовано існуючі підходи до заповнення пропусків, їх переваги та обмеження, а також розглянуто особливості задач класифікації та прогнозування на прикладі медичних та гідрологічних даних як предметних областей застосування запропонованих методів. За результатами проведеного аналізу обґрунтовано доцільність розроблення нових методів імпутування, здатних враховувати зв'язки між ознаками, патерни пропусків та особливості структури даних.

Другий розділ присвячено розробленню алгоритмічного забезпечення інтелектуального імпутування пропусків у даних. У ньому запропоновано нові та вдосконалені методи імпутування `UnifiedClassRegrImputer`, `RegrImputer`, `EntropyImputer`, `HybridRegrEntropyImputer` та `CorrelationImputer`, засновані на використанні різних підходів до обробки даних з урахуванням типу даних, патернів пропусків, а також кореляційних або ентропійних залежностей між ознаками. Для кожного методу наведено UML-діаграму діяльності, що відображає алгоритм його роботи. Також у розділі описано удосконалені методи перетворення якісних ознак на кількісні (`IgnoreNaNLabelEncoder`, `IgnoreNaNFrequentEncoder`), що дозволяють зберігати інформацію про пропуски та виконувати зворотне перетворення.

Третій розділ висвітлює програмну реалізацію запропонованих методів мовою Python. Методи інтегровано в екосистему бібліотеки `scikit-learn`, що забезпечує їх уніфіковане використання в конвеєрах обробки даних. Значну увагу приділено оптимізації програмного коду, зокрема векторизації обчислень та порівняльному аналізу ефективності багатопотокової та багатопроцесної реалізацій ентропійного методу імпутування.

Четвертий розділ містить результати експериментальних досліджень. У розділі проведено порівняльний аналіз точності імпутування, швидкодії та впливу запропонованих методів на якість моделей класифікації й прогнозування часових рядів. Показано, що запропоновані методи забезпечують підвищення точності імпутування та покращують якість моделей порівняно з базовими підходами на реальних наборах даних медичного та гідрологічного моніторингу. Також продемонстровано сумісність розроблених програмних компонентів із конвеєрами обробки даних бібліотеки `scikit-learn`.

У висновках узагальнено результати дослідження, що відображають основні наукові й практичні досягнення роботи.

У додатках представлено список праць здобувача за темою дисертації, лістинги розробленого програмного забезпечення та акт впровадження результатів роботи.

Структура дисертаційної роботи є логічною, послідовною та повністю відповідає поставленій меті й завданням дослідження.

Окремі дискусійні питання і зауваження.

1. Дискусійним є підхід, за якого у підрозділі 1.2, де описано первинний аналіз даних гідрологічного моніторингу, значення одного календарного місяця за різні роки розглядаються як статистична вибірка для застосування критеріїв однорідності. Доцільно було б обґрунтувати коректність такого підходу з урахуванням можливої часової залежності даних та змін статистичних характеристик часових рядів.

2. Опис патернів пропусків на стор. 52 потребує уточнення та узгодження з алгоритмами 2.1–2.3. Зокрема, за буквального тлумачення пункти (а) та (б) можуть сприйматися як взаємовиключні: пункт (а) передбачає відсутність повністю заповнених рядків або стовпців, тоді як пункт (б) – наявність повністю заповнених стовпців. Водночас в описі алгоритмів ключовим етапом є формування підмножини `full_data`, яка складається з повністю заповнених рядків, та не цілком зрозуміло, яким чином враховується патерн (б). У зв'язку з цим доцільно уточнити зміст патернів (а)–(д) та детальніше пояснити, яким чином кожен із них враховується в наведених алгоритмах.

3. У роботі варто уточнити умови, за яких запропонований метод `RegImputer` є коректним для категоріальних ознак, а також пояснити, у яких

випадках коригування прогнозованого регресійною моделлю значення забезпечує коректне відновлення категоріальної змінної.

4. Для запропонованого методу `CorrelationImputer` не цілком зрозуміло, яким чином метод працює у випадку, коли жодна з побудованих регресійних моделей не забезпечує прийнятної якості апроксимації. Також доцільно було б обґрунтувати вибір саме восьми квазілінійних моделей.

5. У підрозділі 3.2 переваги векторизації обчислень продемонстровано на прикладі окремого фрагмента програмного коду. Доцільно було б коротко охарактеризувати, наскільки аналогічні підходи були використані під час реалізації інших розроблених класів імпутерів.

6. У підрозділі 3.3 під час порівняння послідовного, багатопоточного та багатопроцесного підходів доцільно було навести оцінку варіабельності часу виконання (наприклад, стандартне відхилення поряд із середнім), а також для забезпечення коректності порівняння проводити експерименти за однакових масок пропусків і фіксованого випадкового компонента алгоритму імпутації.

7. У підрозділах 3.3 та 4.3 швидкодію реалізованих методів варто було оцінювати окремо для кожного відсотка пропусків, а не лише узагальнено, оскільки обчислювальна складність алгоритмів може істотно залежати від кількості пропущених значень.

8. У підрозділах 4.3 та 4.5 у ході дослідження методів імпутування на даних гідрологічного моніторингу серед базових методів порівняння використано методи `SimpleImputer`, `IterativeImputer`, `KNNImputer`, які не враховують часову структуру даних. У разі нестационарності часового ряду (тренду, структурних зсувів) застосування таких методів без попереднього дослідження стаціонарності може призводити до систематичних похибок імпутації, оскільки заповнене значення не враховує положення пропуску в часі. Тому доцільно було також порівняти запропоновані методи з підходами, що враховують часову структуру даних (наприклад, сезонну чи лінійну інтерполяцію за часом), або обґрунтувати, чому застосування методів, які не враховують часову структуру, є достатнім для поставленої задачі.

9. Для підтвердження переваги запропонованих методів над базовими доцільно було не лише порівнювати середні значення показників (RMSE, ассигасу, recall тощо), а й навести оцінку їх варіабельності та застосувати

статистичні тести (наприклад, тест знакових рангів Вілкоксона або, за виконання відповідних припущень, парний t-тест).

Проте висловлені зауваження не знижують загальної позитивної оцінки дисертаційної роботи, її наукового рівня, а також обґрунтованості основних наукових положень і висновків автора.

Загальний висновок. Дисертаційна робота Земляного Олексія Дмитровича на тему «Розроблення методів та програмного забезпечення інтелектуального імпутування пропусків даних», подана на здобуття ступеня доктора філософії за спеціальністю 121 Інженерія програмного забезпечення, є завершеною науковою працею, присвяченою розв'язанню актуальної наукової задачі, результати якої характеризуються науковою новизною та мають практичне значення. Основні положення, висновки та результати дослідження є обґрунтованими і достовірними. Оформлення дисертації відповідає вимогам, затвердженим наказом Міністерства освіти і науки України від 12.01.2017 № 40. За науковим рівнем, новизною одержаних результатів, практичним значенням та повнотою їх оприлюднення дисертаційна робота відповідає вимогам Постанови Кабінету Міністрів України від 12.01.2022 № 44 (зі змінами) «Про затвердження Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», а її автор Земляний Олексій Дмитрович заслуговує на присудження ступеня доктора філософії за спеціальністю 121 Інженерія програмного забезпечення.

Рецензент

доцент кафедри інженерії програмного
забезпечення та інформаційних технологій
Дніпровського національного університету
імені Олеся Гончара,
кандидат технічних наук, доцент



Ольга МАЦУГА

Підпис канд. техн. наук, доцента Мацуги О.М. засвідчую
Вчений секретар
Дніпровського національного університету
імені Олеся Гончара
кандидат фізико-математичних наук, доцент



Тетяна ХОДАНЕН