УДК 004.89

АДАПТИВНИЙ АЛГОРИТМ ВИБОРУ МЕТОДУ ІМПУТАЦІЇ ПРОПУЩЕНИХ ЗНАЧЕНЬ У ЗАДАЧАХ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Волобуєва Д.С., Губін В.О.
email: daryna.volobuieva@nure.ua, vadim.gubin@nure.ua
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

In this paper, methods for imputing missing values in numerical datasets are analyzed. The study examines mean, median, and kNN imputation, as well as their impact on the distribution and integrity of the data. An adaptive algorithm is proposed, which allows determining the appropriate imputation method for each feature depending on the distribution shape, presence of outliers, and correlations between features. A real numerical dataset with artificially introduced missing values was used for the experiment. The obtained results confirm the necessity of an adaptive approach to imputation in data analysis and machine learning.

У задачах інтелектуального аналізу даних та штучного інтелекту (ШІ) якість вхідних даних впливає на точність побудованих моделей. Некоректні або неповні дані можуть спричиняти зміщення оцінок параметрів, зниження здатності моделей та погіршення прогностичної ефективності. Оскільки системи ШІ базуються на опрацюванні великих масивів знань, їхня результативність визначається повнотою та достовірністю вибірки. У багатьох задачах інтелектуального аналізу даних, пов'язаних з обробкою емпіричних кількісних спостережень, дані можуть бути спотворені через пропуски [1]. Тому, вибір відповідного методу обробки пропущених значень є важливим для отримання якісних результатів аналізу даних.

Найпростішими та найпоширенішими підходами до обробки пропусків є імпутація середнім значенням (mean) або медіаною (median). Дані підходи мають низьку обчислювальну складність та легко реалізуються у табличних процесорах і бібліотеках аналізу даних. Проте вони не враховують структуру даних, взаємозв’язки між ознаками та можуть призводити до надмірної концентрації штучно створених значень. Більш складні методи, зокрема kNN-імпутація, використовують близькі за значеннями ознаки, вимагають додаткових обчислювальних ресурсів та чутливі до вибору параметрів [2]. Водночас такі методи частково враховують локальну структуру даних та приховані закономірності, що особливо важливо для підвищення ефективності систем штучного інтелекту.

У даній роботі пропонується адаптивний алгоритм вибору методу імпутації, який враховує форму розподілу ознак, наявність викидів та

щільність пропущених значень. Замість застосування одного універсального методу до всіх ознак, пропонується здійснювати попередній аналіз даних і обирати метод імпутації індивідуально для кожної числової ознаки [3, 4]. Такий підхід дозволяє зменшити спотворення статистичних характеристик та зберегти структуру даних, що є важливим для подальшого навчання моделей машинного навчання.

Для проведення експерименту використано датасет, що містить числові ознаки, зокрема кількість пропусків занять (absences). З метою оцінки впливу різних методів імпутації на вхідні дані, було штучно згенеровано, з використанням мови Python, пропущені значення у числовій виборці. Зроблено заповнення пропусків трьома різними методами: середнім значенням, медіаною та методом kNN.

Після застосування кожного з методів проведено аналіз розподілу отриманих значень, які порівнювались з початковим розподілом ознаки без пропусків. Для візуалізації результатів побудовані графіки (рис. 1), що дозволяють наочно оцінити ступінь спотворення розподілу після імпутації.

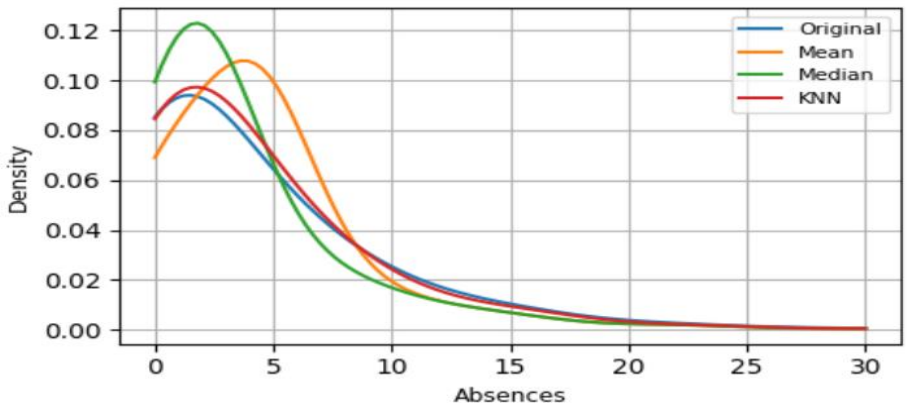


Рисунок 1 – Порівняння методів імпутації для ознаки absences

Отримані результати свідчать, що імпутація середнім значенням призводить до появи аномальної концентрації значень у точці середнього, що зменшує дисперсію ознаки та спотворює форму розподілу. Імпутація медіаною є більш стійкою до викидів, проте занижує ймовірність великих значень. Метод kNN демонструє кращу відповідність початковому розподілу ознаки без пропусків та дозволяє частково зберегти локальну структуру даних. Проте може створювати дробові значення для дискретних ознак та є чутливим до вибору параметра k.

Запропонований алгоритм базується на аналізі статистичних властивостей ознак і реалізується у вигляді набору правил прийняття рішень. Це дозволяє автоматизувати процес вибору стратегії заповнення пропущених значень для кожної числової ознаки окремо:

а) на першому етапі для кожної ознаки обчислюються базові описові статистики та аналізується форма її розподілу з метою виявлення

асиметрії та наявності викидів. Зокрема, визначаються середнє значення, медіана, дисперсія, коефіцієнт асиметрії та показники варіації.

b) на другому етапі оцінюється ступінь статистичного зв'язку ознаки з іншими змінними, що дозволяє визначити потенційну доцільність використання методів, які враховують локальні залежності між об'єктами.

c) на третьому етапі здійснюється вибір методу імпутації згідно з набором правил:

– для ознак із розподілом, близьким до симетричного, та без виражених викидів доцільно застосовувати імпутацію середнім значенням;

– за наявності асиметрії або викидів використовується імпутація медіаною;

– для ознак, що мають значущий статистичний зв'язок з іншими змінними, доцільним є застосування kNN-імпутації.

Проведені дослідження показали, що універсальний метод імпутації не завжди підходить для різних ознак і датасетів. Ефективність методу залежить від типу даних, їх розподілу та структури взаємозв'язків. Запропонований алгоритм вибору методу імпутації забезпечує диференційований підхід до заповнення пропусків у даних та зменшує ризик спотворення статистичних характеристик даних, що є важливим етапом підготовки даних для побудови моделей штучного інтелекту.

Таким чином, застосування адаптивного алгоритму вибору методу імпутації дозволяє більш коректно обробляти пропущені значення та підвищувати ефективність моделей штучного інтелекту. Отримані результати підтверджують доцільність використання адаптивних стратегій імпутації у практичних задачах інтелектуального аналізу даних.

Список використаних джерел:

1. Bodyanskiy Y., Shafronenko A., Klymova I. Adaptive recovery of distorted data based on credibilistic fuzzy clustering approach. CEUR workshop proceedings. 2021. Vol. 2870. P. 6–15. URL: https://ceur-ws.org/Vol-2870/paper2.pdf (дата звернення: 24.02.2026).

2. Niako N., Melgarejo J.D., Maestre, G.E. et al. Effects of missing data imputation methods on univariate blood pressure time series data analysis and forecasting with ARIMA and LSTM. BMC Med Res Methodol 24, 320 (2024). DOI: https://doi.org/10.1186/s12874-024-02448-3.

3. Afkanpour, M., Hosseinzadeh, E. & Tabesh, H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. BMC Med Res Methodol 24, 188 (2024). DOI: https://doi.org/10.1186/s12874-024-02310-6.

4. Ribeiro C., Freitas A.A. A data-driven missing value imputation approach for longitudinal datasets. Artif Intell Rev 54, 6277–6307 (2021). DOI: https://doi.org/10.1007/s10462-021-09963-5.