УДК 004.62

ДВОМЕТРИЧНИЙ ПРОТОКОЛ ОЦІНКИ ТОЧНОСТІ РОЗПОДІЛУ ПРИ ІМПУТАЦІЇ ВІДСУТНІХ ДАНИХ

Гончаренко О.В.
e-mail: oleh.honcharenko@nure.ua
Науковий керівник – к.т.н., проф. Чала Л.Є.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, України

This paper empirically tests the «RMSE illusion» in missing data imputation: the finding that deterministic iterative imputers (MICE, MissForest) minimize reconstruction error while causing Distributional Collapse – measurable degradation of marginal variance and covariance structure. Using the ACI dataset under MNAR conditions, we compare deterministic and stochastic imputers via a dual-metric protocol covering both reconstruction accuracy (RMSE, MAE) and distributional fidelity (Wasserstein distance, VR, KS). Results confirm a fundamental trade-off: deterministic methods achieve low RMSE but fail distributional preservation, while stochastic approaches do the opposite. This motivates a hybrid design combining deterministic initialization with stochastic refinement.

Дана робота описує контрольований бенчмаркінговий експеримент, розроблений для емпіричної перевірки гіпотези про ілюзію RMSE: твердження, що точкові метрики реконструкції, такі як середньоквадратична помилка (RMSE), є недостатніми – і потенційно оманливими – як єдині критерії якості імпутації. Головний аргумент полягає в тому, що детерміновані ітеративні імпутатори систематично оптимізуються для прогнозування умовного середнього, що мінімізує помилку реконструкції, але водночас спричиняє колапс розподілу (Distributional Collapse): вимірюване зменшення маргінальної дисперсії та деградацію структури міжфакторної коваріації в імпутованому наборі даних [1]. Експеримент побудований як спрощене зіставлення двох сімейств імпутаторів за умов контрольованого режиму «невипадкових пропусків» (MNAR) та оцінюється за допомогою системи подвійних метрик, яка незалежно вимірює точність реконструкції та точність збереження розподілу.

Дизайн експерименту ґрунтується на трьох принципах:

– пріоритет надається внутрішній валідності шляхом використання єдиного, добре вивченого повного набору даних, у якому пропуски створюються штучно;

– зберігається екологічна валідність через вибір механізму MNAR, який відображає реальні патерни структурних пропусків.

– експеримент є навмисно мінімалістичним у виборі базової стохастичної моделі: використовуючи простий гауссівський семплер замість

складної генеративної моделі, ізолюється саме механізм, відповідальний за розрив – збереження дисперсії через семплювання – замість того, щоб приписувати це архітектурній складності конкретної моделі.

Експерименти проводяться на наборі даних Adult Census Income, який містить 48842 спостереження та 14 ознак, включаючи безперервні фінансові та категоріальні демографічні змінні. Цей набір обрано з трьох причин: справжні нелінійні залежності між ознаками, загальне визнання набору даних як бенчмарка, фінансова спрямованість з моделюванням «хвостів».

Перед експериментом до усіх безперервних ознак застосовується Z-стандартизація. Записи з наявними пропусками виключаються, що дає повний базовий набір даних ($D_{full} = 45{,}222$), який слугує еталоном.

Стандартна практика використовує механізми MCAR (цілком випадкові) або MAR (випадкові) для зручності обчислень. Однак ці механізми ігнорують значення самих ознак. Режими MNAR, навпаки, вносять селективний зсув: ймовірність пропуску є функцією самого значення [2].

Цільовою змінною для маскування обрано hours-per-week. Індикатор пропуску M_i визначається як імовірнісна функція:

$$P\left(M_i = 1 \middle| \text{age}_i, cp_i\right) = \sigma\left(\beta_1 \cdot \phi\left(\text{age}_i\right) + \beta_2 \cdot \phi(cp_i) + \varepsilon\right), \quad (1)$$

де $\sigma(\cdot)$ – сигмоїдна функція, $\phi(\cdot)$ – Z-стандартизація, $\varepsilon \to \text{U}(0{,}0.1)$ – випадковий шум.

Два детерміновані ітеративні імпутатори реалізовані як «безпечне» сімейство: MICE та MissForest. Обидва оптимізують умовне математичне очікування $E[X_{miss}|X_{obs}]$, що еквівалентно мінімізації втрат L_2 [3].

«Ризиковий» стохастичний імпутатор є навмисно простим: він оцінює умовний розподіл $p(X_{miss}|X_{obs})$ і здійснює вибірку з нього:

$$\widehat{X_{miss}} = \mu_{cond} + \varepsilon, \quad \varepsilon \to N(0, \sigma^2_{cond}). \quad (2)$$

Для оцінювання результатів застосовувався протокол подвійних метрик: метрик точності реконструкції (RMSE, MAE, Відстань Вассерштейна) та метрик дистрибутивної точності (VR, KS, CMD, PCPE, та дельти ексцесу) [4].

Експерименти проводилися за схемою 10-кратної крос-валідації. Аналізувалися чутливість до частки пропусків (20%, 30%, 40%) та сили механізму MNAR (β_1). Візуалізація включає діаграми розсіювання, гістограми накладання розподілів та теплові карти кореляцій.

За результатами експериментальних досліджень вдалося ідентифікувати два різні режими колапсу моди, що існують у взаємодоповнюючих сімействах методів імпутації:

– детерміністичний колапс: MICE та MissForest досягають низького RMSE, але викликають колапс розподілу (VR < 1.0 та висока CMD). Їхні точкові оцінки є точними в очікуванні, але неправильними в розподілі;

– стохастичний колапс: CGS зберігає точність розподілу, але несе штрафи за середньоквадратичне значення помилки через вибірку з хвостів. Його вибірки мають правильний розподіл, але окремо неточні.

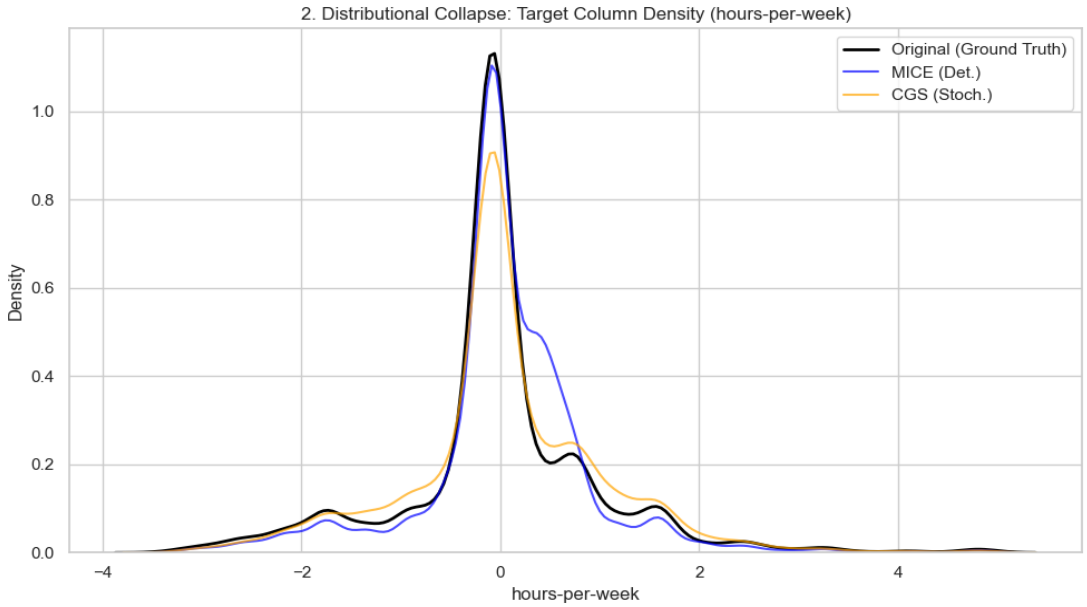

Рисунок 1 – Візуалізація колапсу розподілу

Цей компроміс являє собою доказ існування для взаємодоповнюючого дизайну: архітектури, яка використовує детерміністичну проекцію як теплий старт, а потім стохастичний етап уточнення, який збурює оцінку теплого старту в бік справжнього умовного розподілу. Таким чином, у роботі доведено існування «ілюзії RMSE», що обґрунтовує необхідність розробки детерміновано-стохастичних методів імпутації.

Список використаних джерел:

1. van Buuren S. Flexible Imputation of Missing Data. Chapman and Hall/CRC, 2012. DOI: https://doi.org/10.1201/b11826.

2. Little R. J. A., Rubin D. B. Single Imputation Methods. Statistical Analysis with Missing Data. Hoboken, NJ, USA, 2014. С. 59–74. DOI: https://doi.org/10.1002/9781119013563.ch4.

3. Stekhoven D. J., Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2011. Т. 28, № 1. С. 112–118. DOI: https://doi.org/10.1093/bioinformatics/btr597.

4. Generative adversarial learning for missing data imputation / X. Wang та ін. Neural Computing and Applications. 2024. DOI: https://doi.org/10.1007/s00521-024-10652-x.