

ПІДВИЩЕННЯ ТОЧНОСТІ ВІДНОВЛЕННЯ ПРОПУЩЕНИХ ДАНИХ ЗА ДОПОМОГОЮ ГІБРИДНОГО МЕТОДУ MICE-MLP

Гончаренко О.В., Чала Л.Е.

Кафедра штучного інтелекту,
Харківський національний університет радіоелектроніки,
Україна

E-mail: oleh.honcharenko@nure.ua
larysa.chala@nure.ua

Abstract

The paper presents a comparative study of missing data imputation methods based on machine learning. Five approaches were analyzed: KNN, Random Forest, MICE, MissForest, and the proposed hybrid method MICE-MLP. Experiments on multiple datasets with 20% and 80% of missing data demonstrated that MICE-MLP provides the most stable results with the lowest normalized root mean square error (NRMSE), especially for highly incomplete data.

Якість даних відіграє ключову роль у побудові точних і надійних моделей машинного навчання. Однак у реальних наборах даних часто виникає проблема відсутності значень, коли частина даних є неповною або недоступною через технічні збої або людський фактор. Наявність таких пропусків суттєво впливає на якість аналізу, знижує точність прогнозу та призводить до упереджених висновків.

Оскільки традиційні методи заповнення пропущених значень, такі як заміна середнім або використання регресії, не завжди враховують складні взаємозв'язки в даних, дедалі більшого поширення набувають підходи, засновані на машинному навчанні [1-3]. Проведений аналіз оцінює ефективність використання методів машинного навчання для відновлення пропущених значень у вибірках, де врахування складних взаємозв'язків між змінними є вирішальним для ефективного заповнення пропусків.

Залежно від природи відсутніх даних, виділяють три основні механізми їх відсутності: MCAR, MAR, та MNAR. Така категоризація має вирішальне значення, оскільки вона впливає на похибки при відновленні, що в свою чергу на пряму впливає на надійність обраного методу [4].

Математично механізм відсутності значень у наборі даних визначається залежністю ймовірності пропуску від спостережуваних і прихованих змінних:

$$P(R = 1 | X_{miss}, X_{obs}), \quad (1)$$

де R – матриця, що вказує на відсутність даних;

X_{miss} – відсутні значення;

X_{obs} – наявні спостереження.

Механізм MCAR (Missing Completely At Random) описує ситуацію, коли відсутні дані не залежать ані від спостережуваних даних, ані від пропущених. Тобто пропуски виникають абсолютно випадково, і між фактом їх наявності та доступними даними не існує жодного систематичного зв'язку.

Механізму відсутності MAR (Missing At Random) передбачає, що пропуски залежать лише від спостережуваних даних. При цьому між наявними та пропущеними значеннями існує статистична залежність, яку можна використати для прогнозування відсутніх значень на основі наявної інформації.

Натомість механізм MNAR (Missing Not At Random) описує сценарій, коли відсутність даних не є випадковою і не може бути пояснена спостережуваними змінними. У цьому випадку пропуски мають систематичний зв'язок із факторами, що зумовлюють їх появи, але ці фактори не представлені у доступних даних. Отже отримання неупереджених оцінок можливе лише за умови моделювання самого процесу виникнення пропусків, що робить механізм MNAR особливо складним для задач відновлення відсутніх значень [5].

Типові підходи при боротьбі з пропущеними значеннями поділяються на:

- методи видалення: listwise deletion, pairwise deletion;
- методи одинарної заміни: заміна середнім, медіаною, або модою;
- методи множинної заміни: ітеративне відновлення з використанням ланцюгових рівнянь;
- методи на основі моделей прогнозування: використання регресії, дерев рішень, нейронних мереж тощо.

Одним з класичних методів відновлення пропущених значень є метод MICE, який базується на ідеї послідовного моделювання кожної змінної з пропущеними значеннями на основі решти змінних [6-7]. Цей метод виконує кілька ітерацій заповнення, що дозволяє врахувати взаємозв'язки між змінними та невизначеність при відновленні пропусків єдиним значенням. Однак MICE використовує модель регресії для прогнозування та припускає лінійну залежність між змінними у наборі даних, що не завжди є притаманною реальним наборам даних [8]. Таким чином, існує необхідність розробки гібридних рішень, які поєднують ітеративний підхід MICE з можливостями моделей машинного навчання. Прикладом такого поєднання є метод MissForest, який використовує випадкові ліси для прогнозування пропусків, враховуючи при цьому невизначеність, присутню під час відновлення [9].

У роботі проведено дослідження ефективності існуючих методів відновлення, включаючи KNN, MICE, RandomForest, MissForest, та розробку гібридного методу MICE-MLP, що використовує нейронну мережу для прогнозування пропусків протягом декількох ітерацій заповнення (рис. 1).

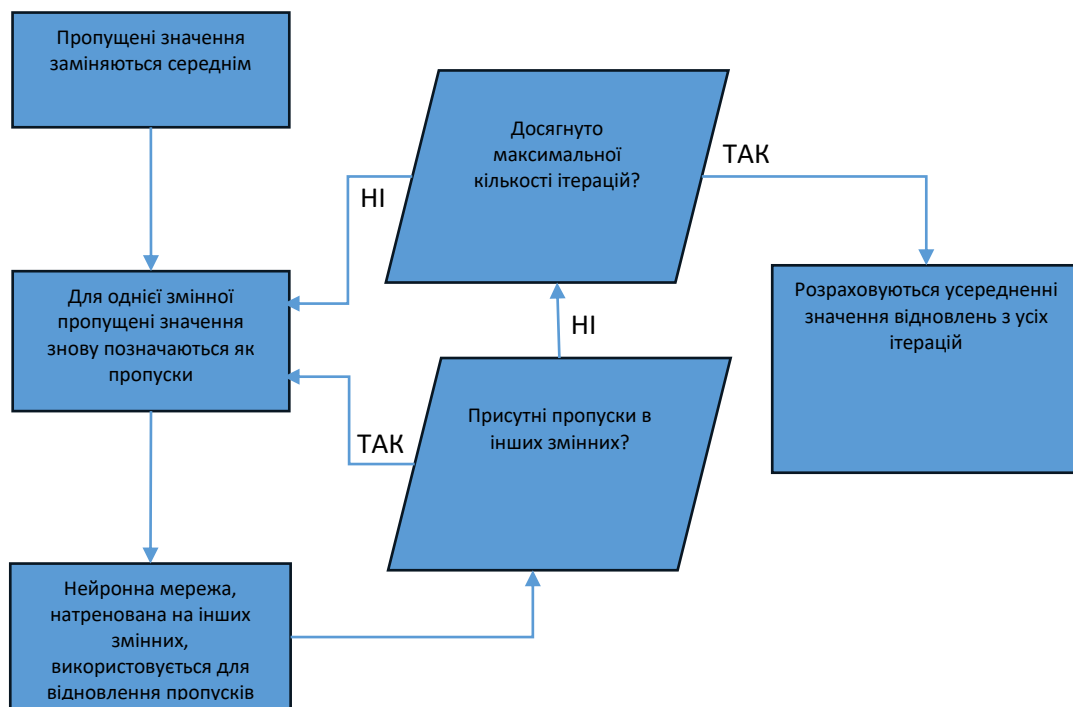


Рис. 1. Принцип роботи запропонованого методу MICE-MLP [10]

Для проведення експериментів було обрано наступні набори даних: Dermatology, BEED, Ozone, SkillCraft, та SUSY. Щоб пропорція пропущених даних повністю контролювалась, оригінальні пропущені значення, присутні у наборах даних, були видалені. Пропорція згенерованих пропусків для кожного набору становила 20% та 80% від загальної кількості прикладів. Враховуючи природу виникнення пропусків у реальних наборах даних, де дані відсутні лише у певній кількості змінних,

генерація пропусків відбувалася з прив'язкою до змінних: випадковим чином обирались k змінних, для яких приклади замінялись на пропуски, де кількість змінних k варіювалась від 1 до $\frac{p}{3}$, де p – загальна кількість змінних у наборі даних. На рис. 2 та рис. 3 наведено динаміку зміни значень похибки відновлення в залежності від застосованих методів та наборів даних.

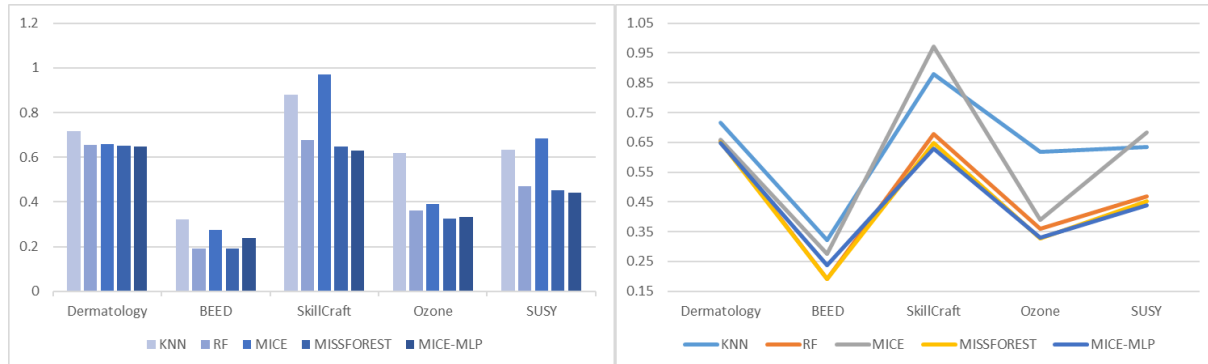


Рис. 2. Похибка методів при відновленні 20% прикладів

При 20% пропущених значень класичні методи KNN, MICE, та RandomForest показали помірні результати. KNN загалом демонстрував найвищі значення NRMSE серед усіх методів, що свідчить про обмеженість метода при відновленні даних зі складними взаємозв'язками між змінними. MICE мав середню стабільність, де його ефективність помітно варіювалась залежно від набору даних, що прямо впливає з припущень методу щодо лінійних зв'язків між змінними. Метод RandomForest мав найкращі результати серед неітераційних методів, особливо на наборах даних BEED та Ozone, де його точність знатно перевищувала KNN та MICE. Метод MICE-MLP продемонстрував одну з найвищих точностей відновлення серед усіх методів, а також конкурентну точність відновлення у порівнянні з методом MissForest. Зокрема, MICE-MLP мав кращі показники NRMSE на наборах даних Dermatology (на 0.38% нижче), SkillCraft (на 3% нижче), та SUSY (на 2.6% нижче), тоді як на наборах даних BEED та Ozone цей метод поступався MissForest (на 2.5% та 1.3% відповідно). У середньому, для такої пропорції пропусків розроблений метод виявився на 0.5-1% точнішим за MissForest, що свідчить про ефективне поєднання нейронної мережі для більш ефективного захоплення складних нелінійних взаємозв'язків.

Таблиця 1. Точність відновлення при 20% пропущених значень

Набір / Метод	KNN	RandomForest	MICE	MissForest	MICE-MLP
Dermatology	0.7151	0.6554	0.6584	0.6505	0.6480
BEED	0.3210	0.1913	0.2747	0.1916	0.2379
SkillCraft	0.8783	0.6783	0.9716	0.6492	0.6298
Ozone	0.6179	0.3598	0.3893	0.3265	0.3309
SUSY	0.6337	0.4648	0.6838	0.4510	0.3894

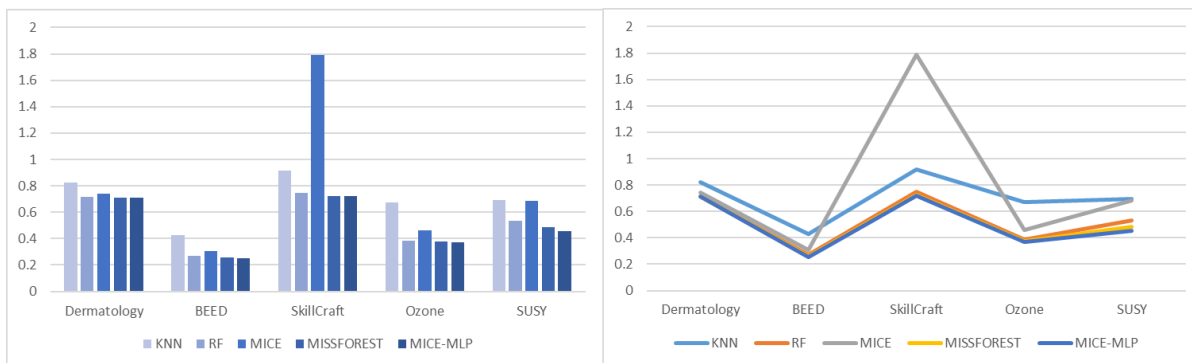


Рис. 3. Похибка методів при відновленні 80% прикладів

За пропорції у 80% пропусків ефективність класичних методів KNN, MICE та RandomForest зменшилась. Зокрема KNN та MICE втратили точність відновлення через принципи їх роботи, які є ненадійними в разі великої кількості пропусків. MICE показав значно вищу похибку при відновленні, особливо на наборі StarCraft, де його лінійна модель не змогла ефективно відновити дані при великій пропорції відсутніх даних. RandomForest зберіг стабільність у роботі і залишився кращим серед трійки класичних методів, дещо поступаючись більш сучасним ітераційним підходам.

Методи MICE-MLP та MissForest продемонстрували найменшу деградацію точності відновленні при збільшенні кількості пропусків. MICE-MLP зберіг конкурентні показники, при цьому випереджаючи MissForest на більшості датасетів. Метод показав кращі результати на наборах BEED (на 2.6%), Ozone (на 2.2%), та SUSY (на 5.8%), при цьому маючи майже однакове значення на наборах Dermatology (на 0.35% гірше) та SkillCraft (на 0.40% краще). Метод був приблизно на 2-3% точнішим за MissForest, при цьому варто зазначити особливу перевагу над класичним MICE, який втрачав точність на складних наборах даних. Порівняно з іншими методами, методи MissForest та MICE-MLP, що використовували ітераційний підхід, показали у середньому на 20% нижчі значення NRMSE, навіть без глибокого налаштування параметрів та за обох пропорцій відсутності.

Таблиця 2. Точність відновлення при 80% пропущених значень

Набір / Метод	KNN	RandomForest	MICE	MissForest	MICE-MLP
Dermatology	0.8236	0.7157	0.7417	0.7133	0.7108
BEED	0.4265	0.2690	0.3069	0.2594	0.2527
SkillCraft	0.9161	0.7491	1.7903	0.7230	0.7198
Ozone	0.6725	0.3850	0.462	0.3772	0.3691
SUSY	0.6943	0.5342	0.6833	0.4839	0.4558

Проведений аналіз показав перспективність використання розробленого гібридного методу, що мав у середньому найвищу точність відновлення порівняно з класичними методами. Його ефективність була особливо помітною у сценарії з високою кількістю пропущених значень, де MICE-MLP перевершив більшість методів за похибкою при відновленні.

Література

1. Ochieng' Odhiambo F. Comparative Study of Various Methods of Handling Missing Data // Mathematical Modelling and Applications. 2020. Vol. 5, No. 2. P. 87.
2. Barzi F. Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies // American Journal of Epidemiology. 2004. Vol. 160, No 1. P. 34–45.
3. Jerez J. M. Missing data imputation using statistical and machine learning methods in a real breast cancer problem // Artificial Intelligence in Medicine. 2010. Vol. 50, No. 2. P. 105–115.
4. Little R. J. A., Rubin D. B. Single Imputation Methods. Statistical Analysis with Missing Data // Hoboken, NJ, USA. 2014. P. 59–74.
5. Buuren S. Flexible Imputation of Missing Data // Chapman and Hall/CRC. 2012.
6. Barnard J., Schenker N., Rubin D. B. Multiple Imputation // International Encyclopedia of the Social & Behavioral Sciences. 2015. P. 88–93.
7. Rubin D. B. Multiple Imputation after 18+ Years // Journal of the American Statistical Association. 1996. Vol. 91, No. 434. P. 473–489.
8. Deng Y. Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data // Scientific Reports. 2016. Vol. 6, No. 1. P. 35–39.
9. Stekhoven D. J., Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data // Bioinformatics. 2011. Vol. 28, No. 1. P. 112–118.
10. Khan S. I., Hoque A. S. M. L. SICE: an improved missing data imputation technique // Journal of Big Data. 2020. Vol. 7, No. 1. P. 115–139.