Developing Robust Methods to Handle Missing Data in Real-world Applications Effectively

Youran Zhou, Sunil Aryal, Mohamed Reda Bouadjenek
School of Information Technology, Deakin University, Geelong, Victoria, Australia

ECML PKDD 2024

PhD Forum-1605

1. Introduction

Why Data Gets Missed?

- Human Error, Systematic Issues, Non-response or Refusal, Data Corruption, Survey Design or Sampling Issues, Conditional Data Collection

Missing data is a common challenge across various data types.

- Most research focuses on the **Missing Completely At Random (MCAR) missing mechanism**.
- This project explores underexamined mechanisms like **Missing At Random (MAR)** and **Missing Not At Random (MNAR).**

2. Background

Missing mechanism Ψ :
How and why data becomes missing.

Missing Mask M:
This mask is used to represent the location of missing data that occur in.

MCAR:
Missingness is random and unrelated to the data.

$$f(\mathrm{M}|\Psi)\ \forall\ X, \Psi$$

MAR:
Missingness is related to observed data.

$$f(\mathrm{M}|X^o, \Psi)\ \forall\ X^m, \Psi$$

MNAR:
Missingness is related to the missing values themselves.

$$f(\mathrm{M}|X^m, \Psi)\ \forall\ X^o, \Psi$$

Table 1. Types of Missing Mechanisms

Gender	Salary			
Complete Dataset		MCAR	MAR	MNAR
F	High	High	High	?
F	High	?	?	?
M	High	?	High	?
F	High	High	?	?
M	High	?	High	?
M	High	High	High	?
M	High	?	High	?
M	Low	Low	Low	Low
F	Low	?	?	Low
M	Low	Low	Low	Low
M	Low	?	Low	Low
F	Low	Low	?	Low
F	Low	Low	?	Low
M	Low	?	Low	Low

3. Research Gap

- **Insufficient** methods for handling MAR & MNAR missing data mechanisms.
- Current methods struggle with **mixed data types** (numerical and categorical).
- Current experiments rely heavily on **synthetic** scenarios.
- Need for more realistic and formulated **missing data generation methods.**

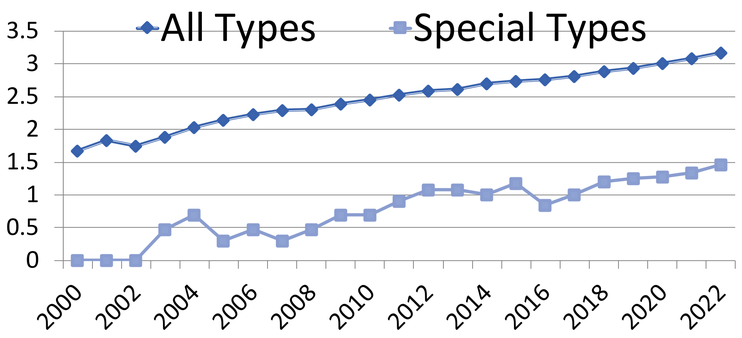


Chart 1. Number of Article occurs in Scopus database via Key words search (In log Scale)

4. Research Aim

Aim: *Developing Methods to Handle different types of missing mechanisms in mixed domain.*

Objective 1: Investigating the effectiveness of existing methods in terms of handling different types of missing mechanisms and data types

Objective 2: Developing robust models for handling diverse types of missing data by investigating and enhancing existing methods to accommodate variations in missing mechanism generation techniques

Objective 3: Extending the novel methods to handle different types of missing mechanisms in categorical and heterogeneous domains.

Objective 4: Extending proposed methods to handle missing modalities in multimodal data.

5. Methodology

- **Diffusion** Based **Imputation** Method
- **Kernel** Based Representation Learning Method with **Heterogeneous Data**
- **Graph Neural Networks** for Handling Special Missing Mechanisms in **Multimodal Data**

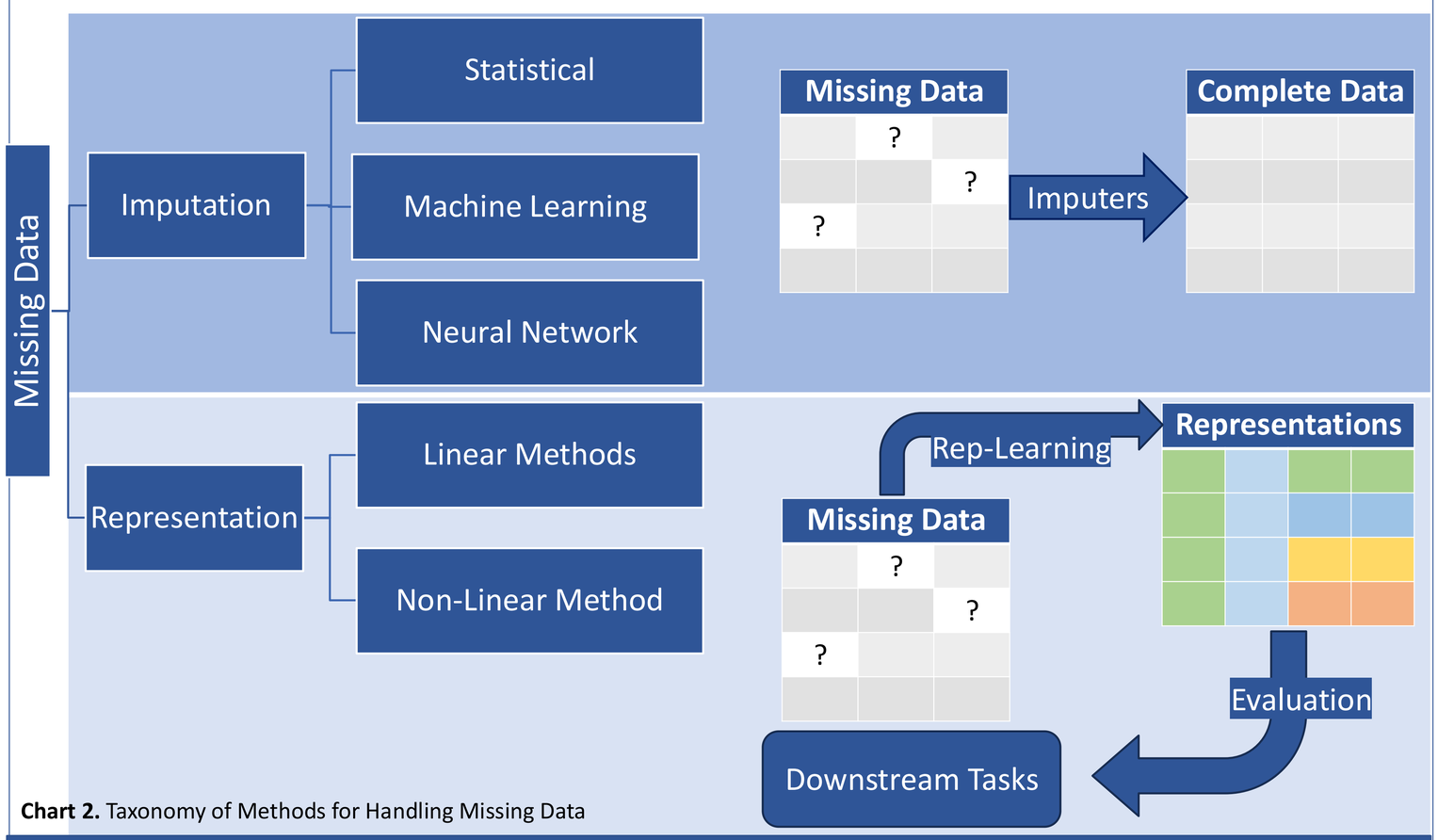
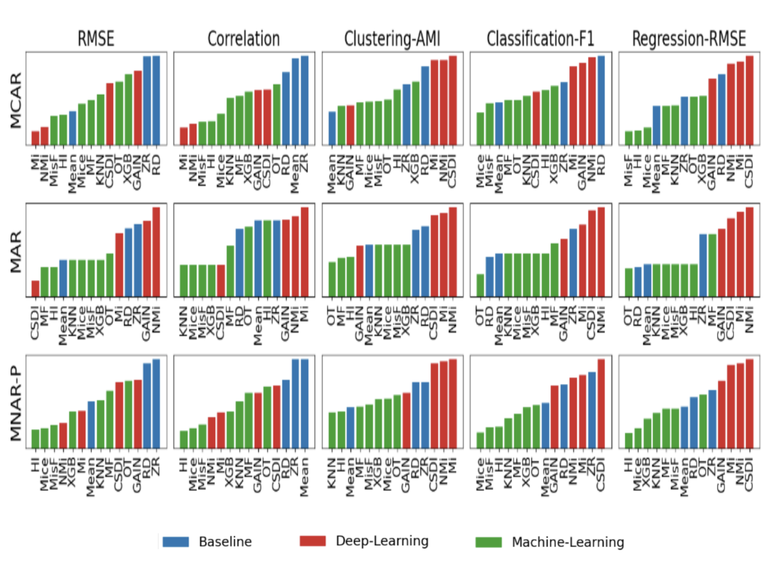


Chart 2. Taxonomy of Methods for Handling Missing Data

6. Current Progress



	Baseline			Machine Learning						Deep Learning			
Ψ	RD	ZR	Mean	KNN	MF	Mice	MisF	XGB	OT	HI	GAIN	Mi	NMi	CSDI
MCAR														
0.3	4.30	4.22	1.86	1.86	1.96	1.59	**1.60**	4.16	2.16	1.66	2.34	1.62	**1.60**	2.22
0.5	4.29	4.22	1.85	2.02	1.99	2.31	1.80	3.98	2.20	1.83	2.64	**1.69**	1.70	2.24
0.7	4.29	4.23	1.86	2.09	2.02	3.08	2.04	3.39	2.20	2.06	3.12	**1.81**	1.92	2.28
MAR														
0.3	4.27	4.05	2.41	2.41	2.21	2.41	2.41	2.41	2.60	2.36	4.08	3.17	74.30	**2.13**
0.5	4.25	4.10	2.53	2.53	2.38	2.53	2.53	2.53	2.59	2.35	4.08	3.31	78.99	**2.11**
0.7	4.39	3.95	2.76	2.76	2.44	2.76	2.76	2.76	2.70	2.48	4.12	3.12	81.68	**2.40**
MNAR-L														
0.3	4.27	4.23	1.88	1.97	1.98	1.99	1.62	3.60	2.15	**1.60**	2.23	1.63	1.64	2.24
0.5	4.28	4.23	1.88	2.06	2.01	2.15	1.82	2.59	2.16	1.78	2.44	**1.69**	1.70	2.22
0.7	4.28	4.20	1.90	2.12	2.04	3.24	2.05	2.59	2.19	2.01	2.89	**1.81**	1.86	3.04

Chart 4. Result from our paper, LHS: Average ranking of different imputation methods w.r.t. RMSE and correlation coefficient of imputed and true value along with performances in three downstream tasks for five different missing mechanisms.
RHS: Average RMSE at different missing parameters/rates.

Contact

Youran Zhou
Deakin University
Email: echo.zhou@deakin.edu.au
LinkedIn: www.linkedin.com/in/youran-zhou

Stay in Touch: Scan to Connect on LinkedIn

Discover More: Scan to Read the Full Paper

References

1. Biessmann, F., Rukat, T., Schmidt, P., Naidu, P., Schelter, S., Taptunov, A., Lange, D., Salinas, D.: Datawig: Missing value imputation for tables. J. Mach. Learn. Res. 20(175), 1–6 (2019)
2. Enders, C.K.: Applied missing data analysis. Guilford Publications (2022)
3. Ipsen, N.B., Mattei, P.A., Frellsen, J.: not-miwae: Deep generative modelling with missing not at random data. arXiv preprint arXiv:2006.12871 (2020)
4. Mattei, P.A., Frellsen, J.: Miwae: Deep generative modelling and imputation of incomplete data sets. In: International conference on machine learning. pp. 4413–4423. PMLR (2019)
5. Nazabal, A., Olmos, P.M., Ghahramani, Z., Valera, I.: Handling incomplete heterogeneous data using vaes. Pattern Recognition 107, 107501 (2020)
6. Zheng, S., Charoenphakdee, N.: Diffusion models for missing value imputation in tabular data (2023)
7. Zhou, Y., Aryal, S., Bouadjenek, M.R.: Review for handling missing data with special missing mechanism (2024)
8. Zhou, Y., Bouadjenek, M. R.& Aryal, S. (2024). **Missing Data Imputation: Do Advanced ML/DL Techniques Outperform Traditional Approaches?** Manuscript accepted for publication in the proceedings of ECML PKDD 2024.