Unsupervised Anomaly Detection in The Presence of Missing Values

Feng Xiao[1] Jicong Fan[1,2]

[1]The Chinese University of Hong Kong, Shenzhen, China
[2]Shenzhen Research Institute of Big Data, Shenzhen, China

Main Contributions

- Study the imputation bias problem of the “impute-then-detect” pipeline and quantitatively evaluate their detection performance.
- Propose ImAD, an end-to-end unsupervised anomaly detection method in the presence of missing value.
- Provide theoretical guarantees for ImAD, proving that it can correctly detect anomalies with high probability.

Motivation

- Unsupervised anomaly detection in incomplete data remains largely unaddressed
  - Anomaly detection methods typically require fully observed data for model training and inference and cannot handle incomplete data.
  - Missing data problem is pervasive in science and engineering, leading to challenges in many important detection tasks, such as abnormal user detection in recommendation systems or novelty cell detection in bioinformatics, where the missing rates can be higher than 30% or even 80%.
  - “Impute-then-detect” methods easily suffer from imputation bias from normal data and are incline to make incomplete sample “normal”.

Proposed Method

- ImAD couples data imputation with anomaly detection to a unified framework that consists of Imputer, Projector and Reconstructor.

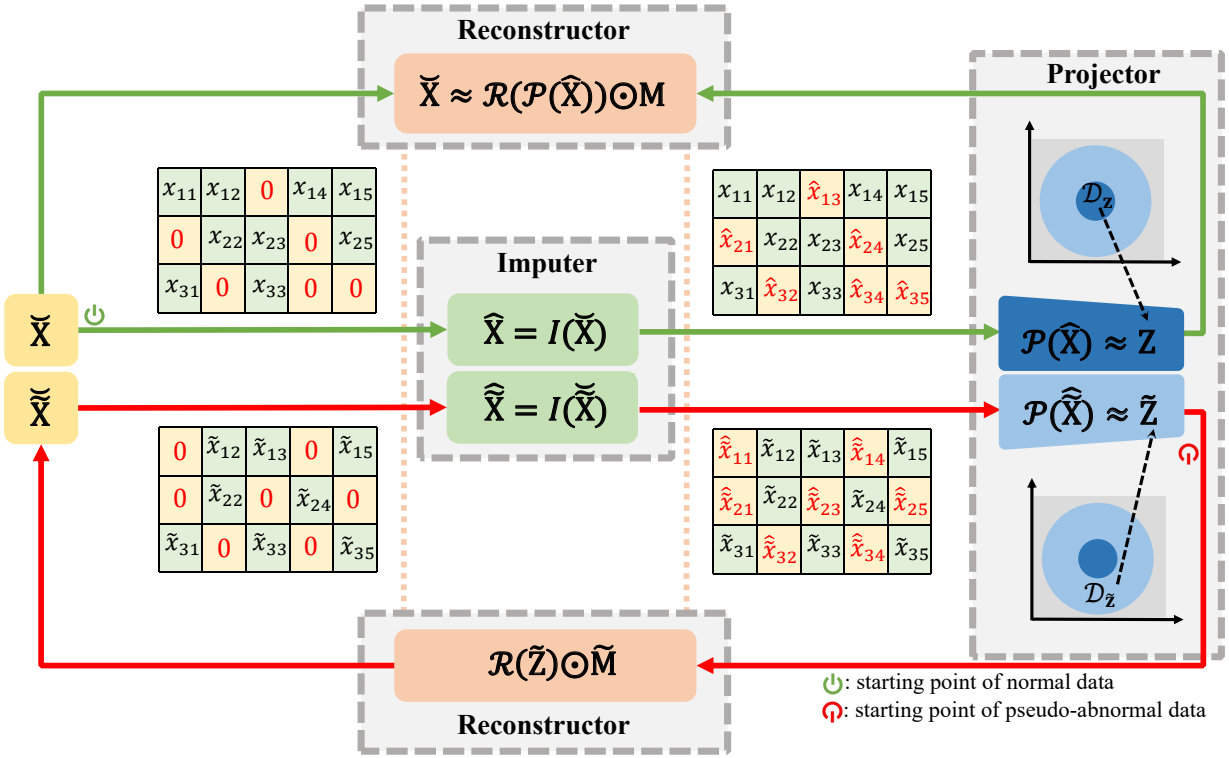


Figure: ImAD framework. $\breve{\mathbf{X}}$ and $\breve{\breve{\mathbf{X}}}$ denote the normal and pseudo-abnormal data with missing values, respectively, while $\hat{\mathbf{X}}$ and $\hat{\hat{\mathbf{X}}}$ are the corresponding imputed data. The Reconstructors share parameters.

Proposed Method

- Imputer aims to recover the missing values of normal and abnormal data.
- Projector aims to project normal and abnormal data into different regions in $\mathcal{Z}$ and makes the model discriminative.
- Reconstructor ensures that the projected data are meaningful in $\mathcal{Z}$.

Optimization problem

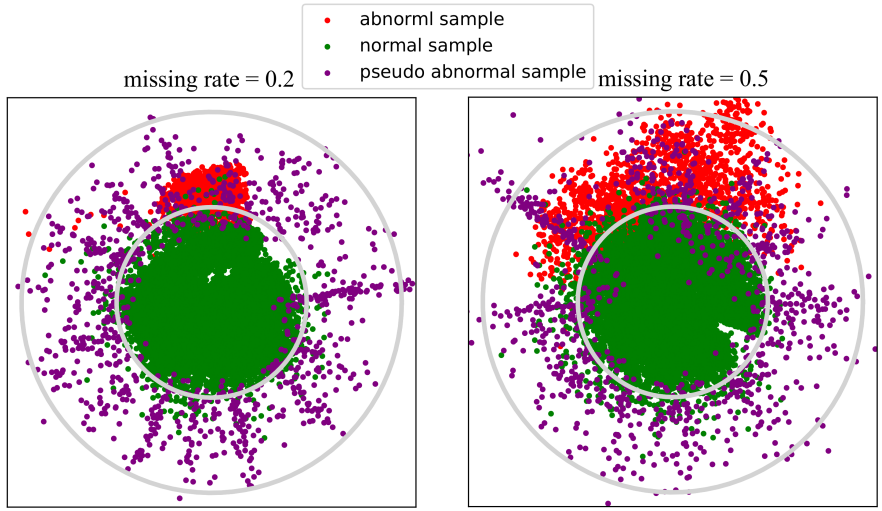
- Optimization Objective of ImAD:

$$\underset{\psi,\theta,\phi}{\text{minimize}} \underbrace{\text{Sinkhorn}(f_\theta(h_\psi(\check{\mathbf{X}})), \mathbf{Z}) + \alpha\|\tilde{\mathbf{Z}} - f_\theta(h_\psi(g_\phi(\tilde{\mathbf{Z}}) \odot \tilde{\mathbf{M}}))\|_F^2}_{\mathcal{L}^{(\text{AD})}}$$
$$+ \underbrace{\beta\|([\check{\mathbf{X}}; \check{\tilde{\mathbf{X}}}] - h_\psi([\check{\mathbf{X}}; \check{\tilde{\mathbf{X}}}])) \odot [\mathbf{M}; \tilde{\mathbf{M}}]\|_F^2}_{\mathcal{L}^{(\text{DI})}}$$
$$+ \underbrace{\lambda\|(\check{\mathbf{X}} - g_\phi(f_\theta(h_\psi(\check{\mathbf{X}})))) \odot \mathbf{M}\|_F^2}_{\mathcal{L}^{(\text{RE})}}$$

- $\mathcal{L}^{(\text{AD})}$ denotes the anomaly detection loss.
- $\mathcal{L}^{(\text{DI})}$ denotes the data imputation loss.
- $\mathcal{L}^{(\text{RE})}$ denotes the reconstruction loss.

Experiments on Synthetic Incomplete Data (MCAR)

- 2-D Visualization of the generated pseudo-abnormal samples.



- Numerical performance evaluation.

DI Methods	AD Methods	Arrhythmia				Speech			
		AUROC(%)		AUPRC(%)		AUROC(%)		AUPRC(%)	
		mr = 0.2	mr = 0.5	mr = 0.2	mr = 0.5	mr = 0.2	mr = 0.5	mr = 0.2	mr = 0.5
MissForest	I-Forest	80.72	81.54	77.91	77.95	28.58	29.09	36.83	37.29
	Deep SVDD	72.63	75.80	70.94	77.39	60.37	40.14	58.93	42.08
	NeutraL AD	47.38	44.30	50.87	50.12	56.51	54.11	55.44	52.26
	DPAD	80.79	**82.64**	80.35	83.30	44.81	43.32	48.57	47.63
GAIN	I-Forest	77.19	76.29	76.40	76.29	29.33	29.23	39.92	40.04
	Deep SVDD	57.14	48.86	59.35	54.03	54.95	46.54	54.38	47.54
	NeutraL AD	37.96	33.98	42.57	42.35	56.80	57.24	54.76	55.05
	DPAD	79.07	80.11	76.48	79.67	41.91	44.95	46.46	49.69
ImAD (Ours)		**82.24**	81.76	**83.74**	**83.37**	**61.94**	**58.66**	**60.43**	**58.13**

Experiments on Real Incomplete Data

- Statisics of real incomplete datasets.

Dataset	Field	Dimension	Missing Samples Rate	Missing Entries Rate
Titanic	pattern recognition	9	79.46%	10.79%
MovieLens1M	recommendation system	498	100%	82.41%
Bladder	cell analysis	23,341	100%	86.93%
Seq2-Heart	cell analysis	23,341	100%	88.51%

- Numerical performance evaluation.

DI Methods	AD Methods	Titanic		MovieLens1M		Bladder		Seq2-Heart	
		AUROC(%)	AUPRC(%)	AUROC(%)	AUPRC(%)	AUROC(%)	AUPRC(%)	AUROC(%)	AUPRC(%)
MissForest	I-Forest	79.72	78.50	36.34	41.45	44.53	46.84	64.58	58.56
	Deep SVDD	60.46	60.78	56.04	53.79	95.53	97.04	94.29	92.30
	NeutraL AD	54.63	52.13	57.14	55.07	66.41	68.01	91.80	90.87
	DPAD	68.18	70.01	47.50	48.49	96.96	97.29	78.03	74.67
GAIN	I-Forest	79.46	78.69	64.84	62.63	45.77	47.62	64.62	58.82
	Deep SVDD	70.59	66.43	58.99	56.68	95.43	96.78	93.93	91.54
	NeutraL AD	53.71	51.55	50.72	51.47	65.30	65.68	91.48	90.79
	DPAD	78.12	77.41	59.98	58.98	96.89	97.25	74.99	73.16
ImAD (Ours)		**82.09**	**81.39**	**66.32**	**65.34**	**100**	**100**	**96.62**	**96.40**

The End

More numerical results as well as the theoretical guarantees can be found in our paper.

Thanks for your attention!