Adaptive Data Augmentation for Supervised Learning over Missing Data

Tongyu Liu, Ju Fan, Yinqing Luo, Xiaoyong Du

中國人民大學
RENMIN UNIVERSITY OF CHINA

Nan Tang

Guoliang Li

清華大學
Tsinghua University

Missing Data is Everywhere

- The raw data we collect from the real world always contain missing values.

Human errors

Sensor mistakes

…. ….

- How to handle the missing values?
  - Dropping the records containing missingness.
  - **Using Imputation methods to complete data.**
    - **It is hard to obtain the ground-truth via value imputation**

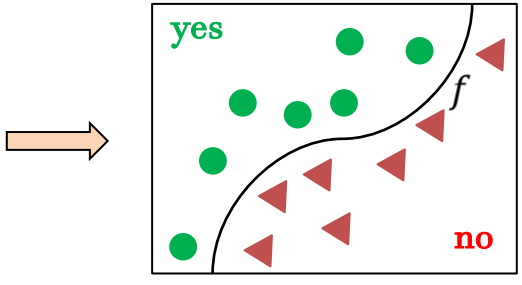
Noise Shift in Source and Target Datasets

- Separately impute source and target data might cause an even bigger divergence on data distributions.

Source Data

labels

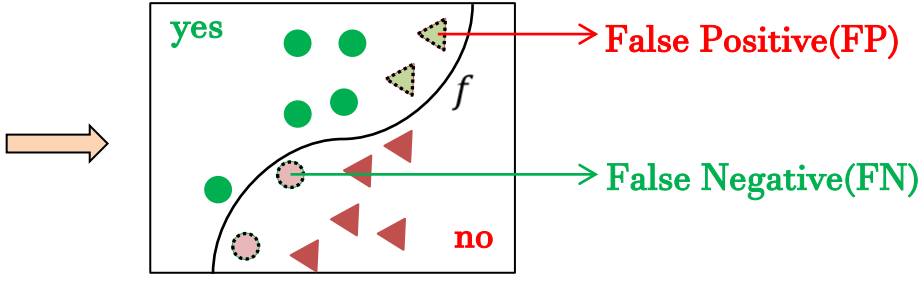
age	cholesterol	glucose	smoke	alcohol	cardio
35	2	1	1	0	no
50	3	3	NA	0	yes
65	2	3	1	NA	yes



Target Data

no labels

age	cholesterol	glucose	smoke	alcohol	cardio
25	1	NA	0	1	?
37	NA	3	0	0	?
70	3	2	1	1	?



If we cannot repair them to be correct, can we repair them to be "similar"?

Adaptive Data Augmentation

Source Data

age	cholesterol	glucose	smoke	alcohol	cardio
35	2	NA	1	0	no
50	3	3	0	0	yes
65	2	3	1	1	yes

① Target Mask Generation

Target Pattern

Target Data

age	cholesterol	glucose	smoke	alcohol	cardio
25	1	NA	0	1	?
37	NA	3	0	0	?
70	3	2	1	1	?

② Source Data Adaptation

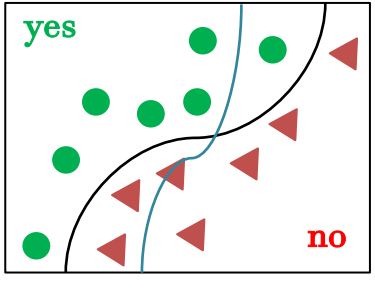
age	cholesterol	glucose	smoke	alcohol	cardio
35	2	2	1	0	no
50	3	3	0	0	yes
65	2	3	1	1	yes

③ Simple Imputation

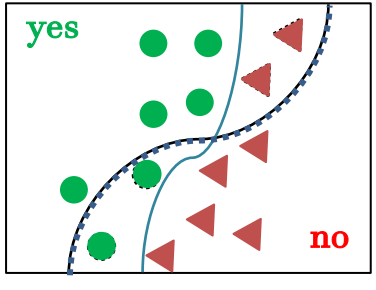
age	cholesterol	glucose	smoke	alcohol	cardio
25	1	2	0	1	?
37	2	3	0	0	?
70	3	2	1	1	?

④ Retrain the model

Source Data

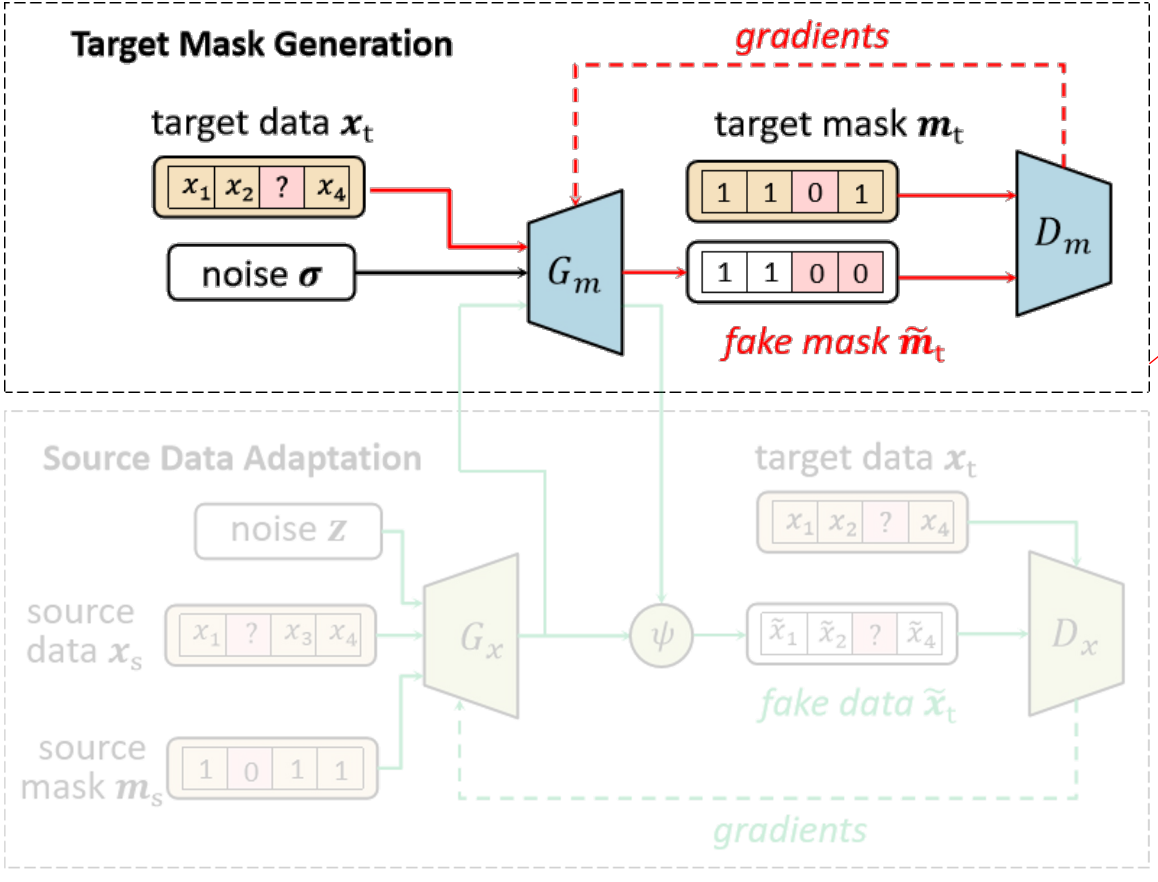


Target Data



⑤ Apply the new model

Target Mask Generation



We utilize Conditional GAN to learn the p(mask|observed data)

$$\mathcal{L}_m(D_m, G_m) = \mathbb{E}_{(x_t, m_t)\sim\mathcal{U}_t}[D_m(m_t, x_t)] - \mathbb{E}_{\sigma\sim p(\sigma),(x_t,m_t)\sim\mathcal{U}_t}[D_m(G_m(\sigma, x_t), x_t)]$$

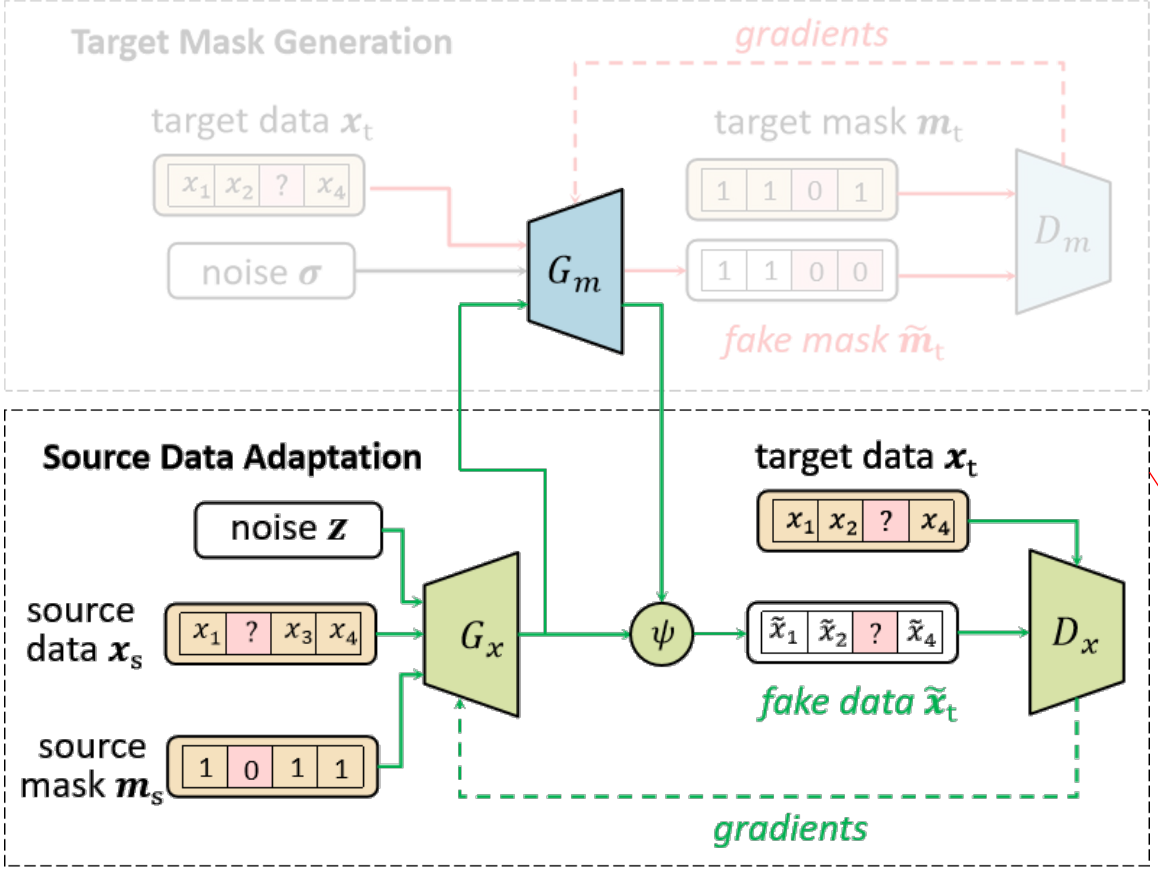
Target Data

age	cholesterol	glucose	smoke	alcohol
25	1	NA	0	1
37	NA	3	0	0
70	3	2	1	1

Mask Matrix

age	cholesterol	glucose	smoke	alcohol
1	1	0	1	1
1	0	1	1	1
1	1	1	1	1

Source Data Adaptation



Adapted Source Data

age	cholesterol	glucose	smoke	alcohol
35	2	NA	1	0
50	3	3	0	0
65	2	3	1	1

Be "Similar"

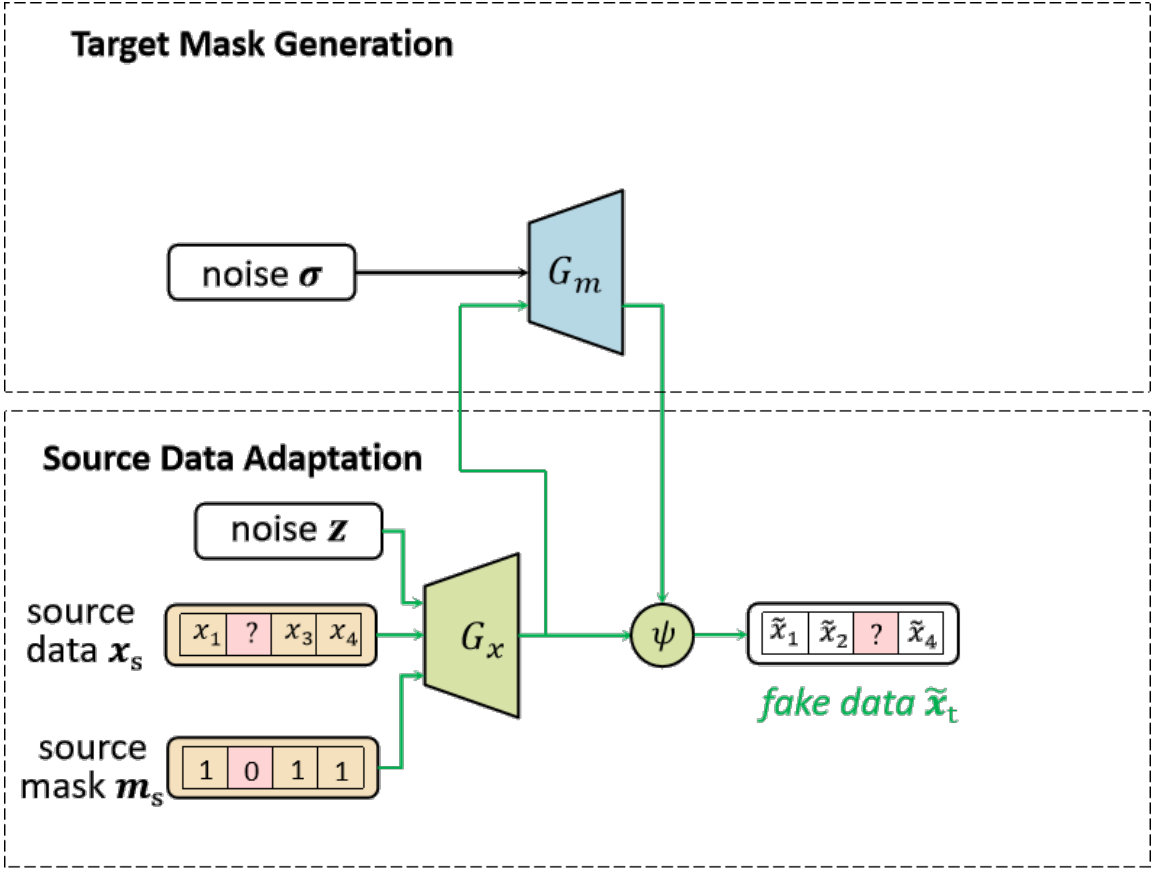
age	cholesterol	glucose	smoke	alcohol
25	1	NA	0	1
37	NA	3	0	0
70	3	2	1	1

Target Data

We then use another GAN to learn the p(target data)

$$\mathcal{L}_x(D_x, G_x, G_m) = \mathbb{E}_{(x_t, m_t)\sim\mathcal{U}_t}[D_x(x_t)]$$
$$-\mathbb{E}_{\sigma\sim p(\sigma), z\sim p(z), (x_s, m_s)\sim\mathcal{D}_s}[D_x(\psi(\tilde{x}, G_m(\sigma, \tilde{x})))], \tilde{x} = G_x(z, x_s, m_s)$$

Adaptive Data Augmentation



Finally, we can transform a source data to the target data by using the mask generator G_m and data generator G_x.

Experiment

- Datasets

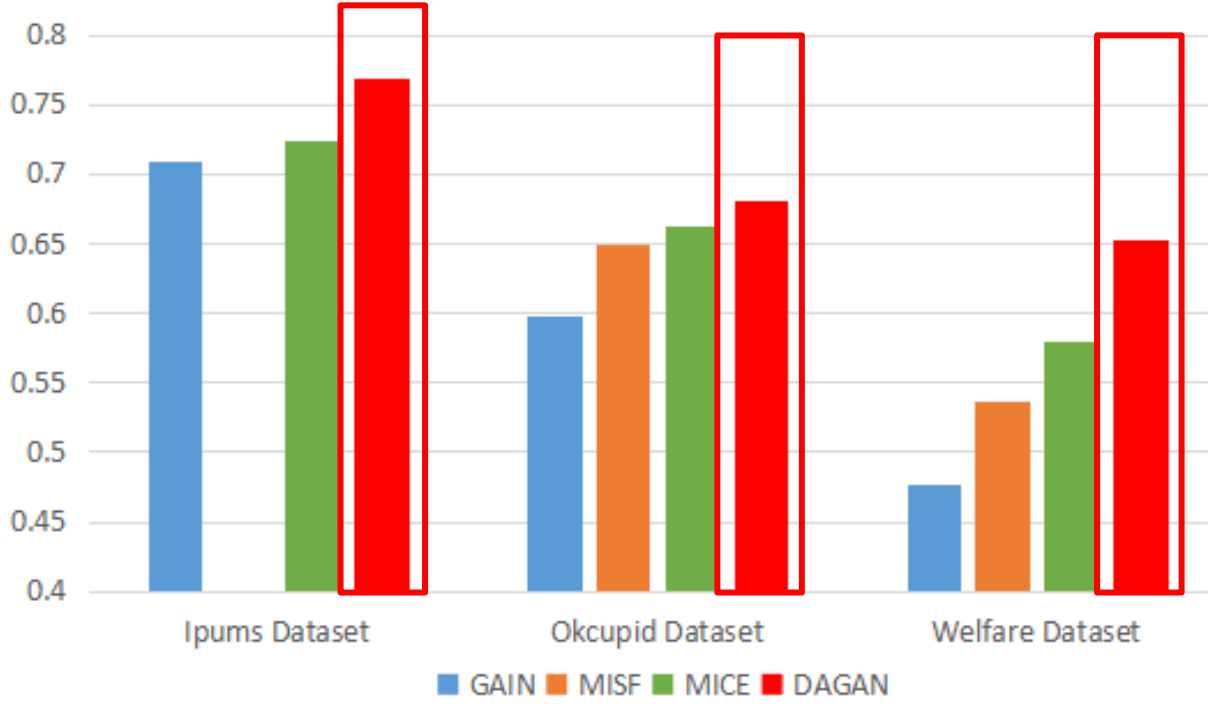
	Dataset	Area	#Rec	#N	#C	#L	Label Distribution
Real Missing	Ipums	Social	16329	16	43	7	65:9:11:5:6:3:1
	Okcupid	Social	50789	3	13	3	7 : 2 : 1
	Welfare	Financial	20309	5	0	2	1.0 : 1.2
Synthetic Missing	EyeState	Health	14977	14	0	2	1.0 : 1.3
	Adult	Social	13567	6	8	2	1.0 : 3.0

- Baselines:
  - MICE: Multiple Imputation with Chained Equations.
  - MISF: MissForest.
  - GAIN: Generative Adversarial Imputation Nets.

Experiment

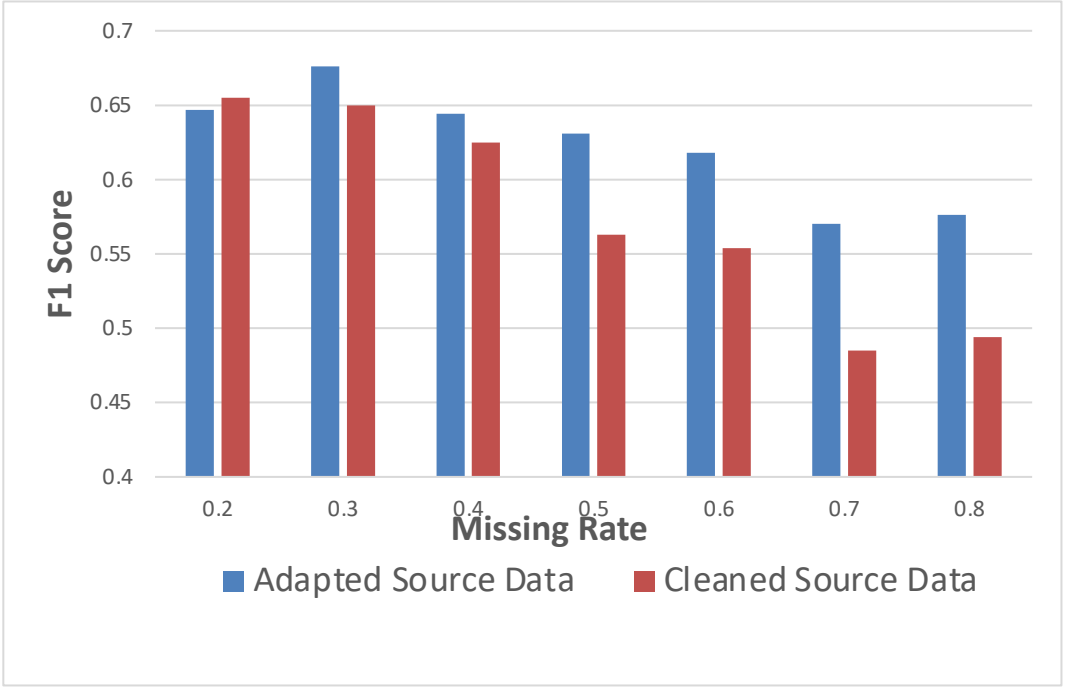
- Evaluation Framework
  - ***Evaluation Task:***
    - Classification.
  - ***Metrics:***
    - F1 score of model.
  - ***For Imputation Methods:***
    - Train the model on **imputed source data**, then apply it to the **imputed target data**.
  - ***For DAGAN:***
    - Train the model on **adapted source data**, then directly apply it to the **target data.**

Evaluation on Real-World Datasets.



* MISF fails on the Ipums dataset because its R-based implementation cannot handle categorical attributes with more than 53 categories

Evaluating Effect of Data Adaptation

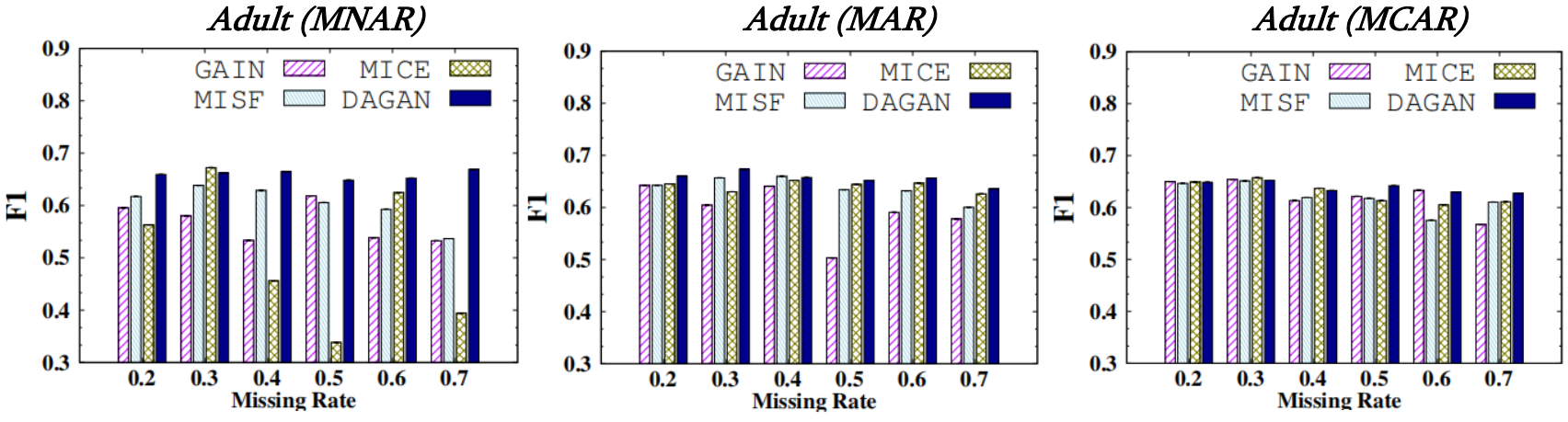


Adult Dataset

Repairing source and target to be "similar" is beneficial for ML training.

Robustness Evaluation

- We consider three different settings of missing value injection.
  - MCAR (Missing Completely at Random)
  - MAR (Missing at Random)
  - MNAR (Missing Not at Random)



DAGAN is robust for different missing data patterns.

Conclusion

- We propose DAGAN to adapt the source data to better serve the prediction on the unseen target data.
- DAGAN performs well over different missing patterns.

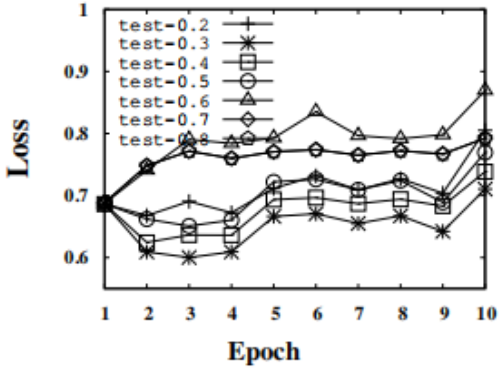
Missing	Dataset-Setting	GAIN	MISF	MICE	DAGAN
MNAR	EyeState-Overlap	-	+	- -	++
	Adult-Overlap	+	-	- -	++
	EyeState-NoOverlap	-	- -	+	++
	Adult-NoOverlap	-	- -	+	++
MAR	EyeState-Overlap	- -	+	-	++
	Adult-Overlap	- -	-	+	++
	EyeState-NoOverlap	-	- -	+	++
	Adult-NoOverlap	- -	-	++	+
MCAR	EyeState-Overlap	- -	++	-	+
	Adult-Overlap	-	- -	+	++
	EyeState-NoOverlap	- -	++	+	-
	Adult-NoOverlap	- -	++	-	+

Average F1 across missing rates under different datasets and missing patterns

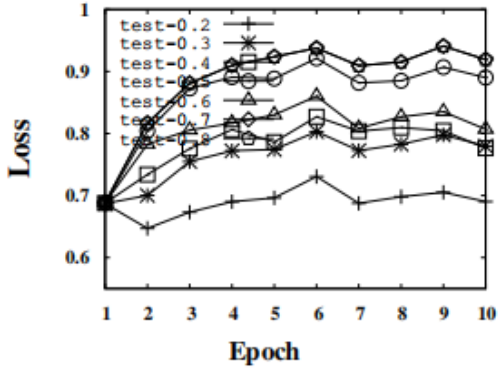
Thank you!

Github repository: https://github.com/ruc-datalab/dagan

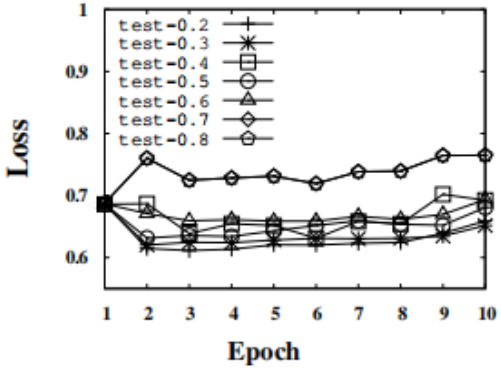
Multi-ADA



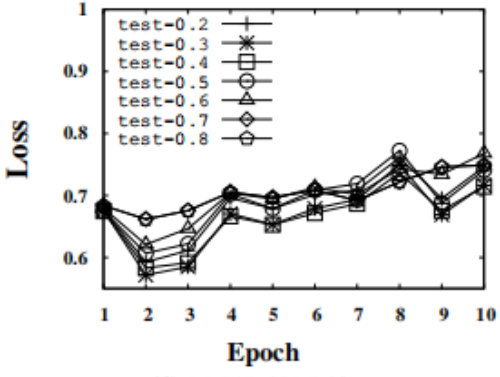
(a) MNAR (ERM)



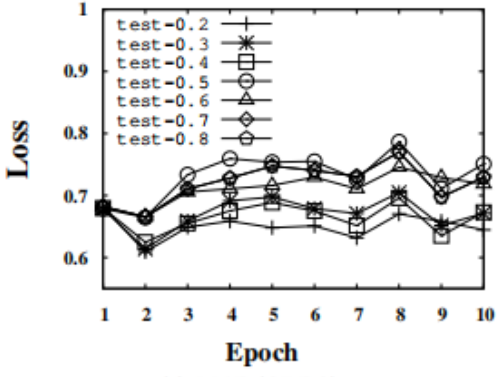
(b) MAR (ERM)



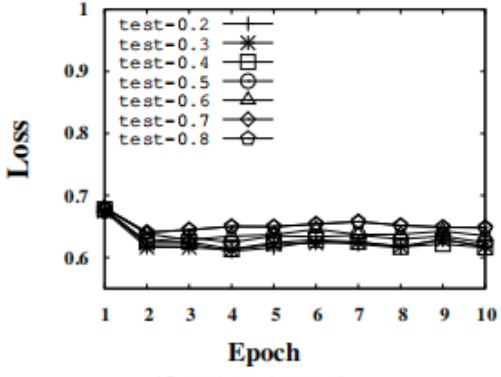
(c) MCAR (ERM)



(d) MNAR (GDRO)



(e) MAR (GDRO)



(f) MCAR (GDRO)