GAIN : Missing Data Imputation using Generative Adversarial Nets

Jinsung Yoon, James Jordon, Mihaela van der Schaar

University of California, University of Oxford, Alan Turing Institute

ICML 2018

2021.02.25

최영제

0 SIL Index

1. Introduction
2. Related works
3. Proposed methodology
4. Results

결측치의 가정

- 결측치를 내포한 데이터의 경우 왜곡된 결과를 도출할 수 있어 분석에 앞서 이를 정제해주는 과정이 필요함
- 결측치는 1) MCAR, 2) MNAR, 3) MAR 3가지 가정으로 구분할 수 있음

1) MCAR (Missing Completely at Random)
   - **결측이 완전히 무작위로 발생한 경우**이며 결측된 이유가 데이터의 어떤 변수와도 관련이 없는 경우
   - 예 : 특정 기간 동안 연구실 사람들의 체온을 측정한 엑셀 파일이 삭제되어 결측이 발생
2) MNAR (Missing Not at Random)
   - 결측치가 특정 의미를 내포하며 그 의미가 **결측된 변수와 관련이 있는 경우**
   - 예 : 고열이 발생한 연구원들은 측정을 거부하여 체온 변수에 결측이 발생
3) MAR (Missing at Random)
   - 결측치가 특정 의미를 내포하며 그 의미가 **데이터의 다른 변수와 관련이 있는 경우**
   - 예 : 체온 측정 시 회식에 참석한 연구원들은 측정이 불가하여 체온 변수에 결측이 발생

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

결측치 보완 방법

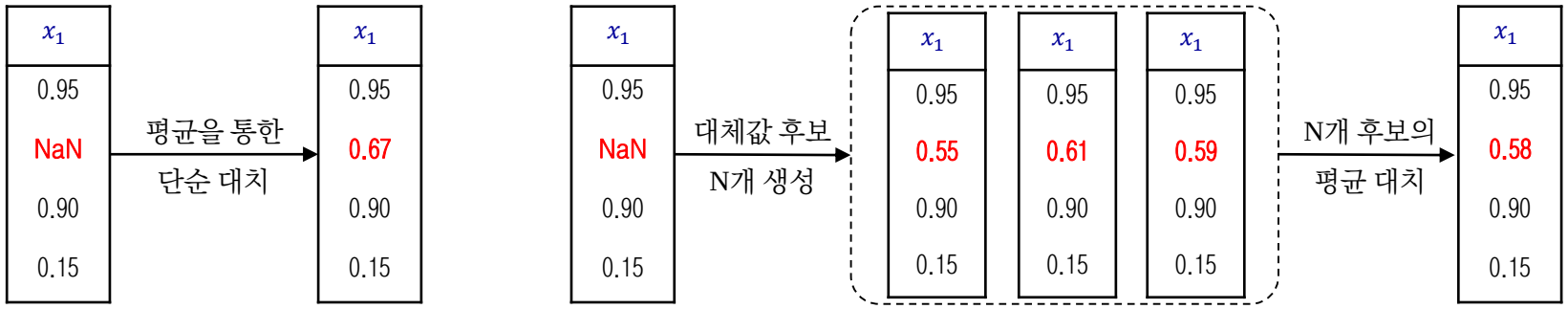
- 결측치를 다루는 방법에는 삭제(deletion)와 대치(imputation)가 존재
- 삭제란 결측치를 포함하는 학습 데이터를 모두 제거하는 방법을 말하며 MCAR에 해당하거나 결측 데이터가 많지 않은 경우 사용함
- 비교적 간단하며 computational cost를 줄일 수 있다는 장점이 있으나 결측 발생의 원인이 MNAR, MAR일 경우 분포를 왜곡
- 대치란 결측치를 임의의 값으로 바꾸어 넣는 것을 말하며 일반적으로 결측치 보완이라 함은 대치를 뜻함
- 기본적인 통계량 대치를 비롯하여 회귀 모형, 머신러닝 모델을 활용하는 등 다양한 방법이 존재
- 결측치 발생 비율에 따른 적합한 보완 방법은 Hair et al. (2006)를 참고하면 아래와 같음

결측치 비율	결측치 처리 방법론
10% 미만	삭제 혹은 어떠한 방법의 대치를 사용해도 좋음
10%~20%	Hot-deck, regression 혹은 model based method
20% 이상	Regression 혹은 model based method

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). Multivariate data analysis (6th ed.). pp. 49-73. Upper Saddle River, NJ: Pearson Educations, Inc.

단순 대치와 다중 대치

- 결측치를 대치하는 방법론은 단순 대치(single imputation)와 다중 대치(multiple imputation)로 구분
- 단순 대치란 결측치를 단일 값으로 보완하는 것을 말함
- 그러나 단일 값으로 결측치를 대체하기에 불확실한 추정값에 대한 과도한 신뢰도를 지닐 수 있음 (표준 오차 과소 평가)
- 다중 대치란 통계적 모형 및 머신러닝 모델 등을 이용하여 결측치가 처리된 n개의 데이터 셋을 생성 후 이를 하나로 종합하여 결측치를 대체
- 단순 대치에 비해서 신뢰성 있는 추론이 가능하며 결측치의 신뢰구간 또한 추정할 수 있다는 장점이 있음



단순 대치 다중 대치

2 Related works

MissForest

- MissForest는 single imputation 기반의 결측치 대치 방법으로 random forest의 예측치를 결측치 대치 값으로 활용
- 총 4가지 step이 존재함
- Step 1) 평균값으로 결측치를 임시 대체한 예

MissForest 알고리즘 전개	
Step 1	평균이나 다른 대치 방법을 사용하여 결측치를 채움
Step 2	결측치 비율이 가장 낮은 변수부터 Random Forest를 사용하여 예측을 진행함. 이 때 학습 데이터는 결측이 발생하지 않은 샘플이며 이를 토대로 결측치에 대해 예측을 진행함
Step 3	모든 변수에 대해서 예측이 종료되면 Step2의 결측이 보완된 데이터를 기반으로 다시 변수 별 예측을 진행함
Step 4	추정된 결측치의 변화량이 특정 범위 이내로 수렴할 때까지 Step2와 Step3을 반복. 수렴 조건을 만족하면 이를 최종 데이터로써 활용함

x_1	x_2	x_3
0.95	1.24	1.46
NaN	0.57	NaN
0.90	NaN	1.28
0.15	0.42	NaN

원본 결측 데이터

x_1	x_2	x_3
0.95	1.24	1.46
0.67	0.57	**1.36**
0.90	**0.74**	1.28
0.15	0.42	**1.36**

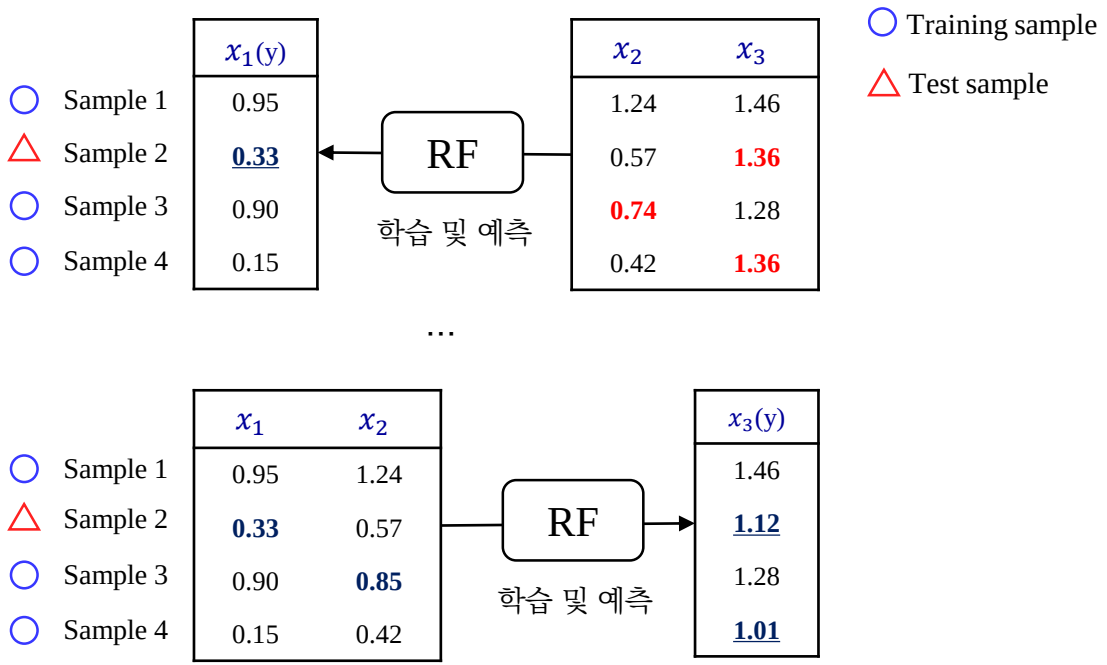
평균 대치 데이터

2 Related works

MissForest

- MissForest는 single imputation 기반의 결측치 대치 방법으로 random forest의 예측치를 결측치 대치 값으로 활용
- 총 4가지 step이 존재함
- Step 2) 학습 데이터를 통해 Random Forest를 학습하고 변수 별 결측치를 예측하는 과정

MissForest 알고리즘 전개	
Step 1	평균이나 다른 대치 방법을 사용하여 결측치를 채움
Step 2	결측치 비율이 가장 낮은 변수부터 Random Forest를 사용하여 예측을 진행함. 이 때 학습 데이터는 결측이 발생하지 않은 샘플이며 이를 토대로 결측치에 대해 예측을 진행함
Step 3	모든 변수에 대해서 예측이 종료되면 Step2의 결측이 보완된 데이터를 기반으로 다시 변수 별 예측을 진행함
Step 4	추정된 결측치의 변화량이 특정 범위 이내로 수렴할 때까지 Step2와 Step3을 반복. 수렴 조건을 만족하면 이를 최종 데이터로써 활용함

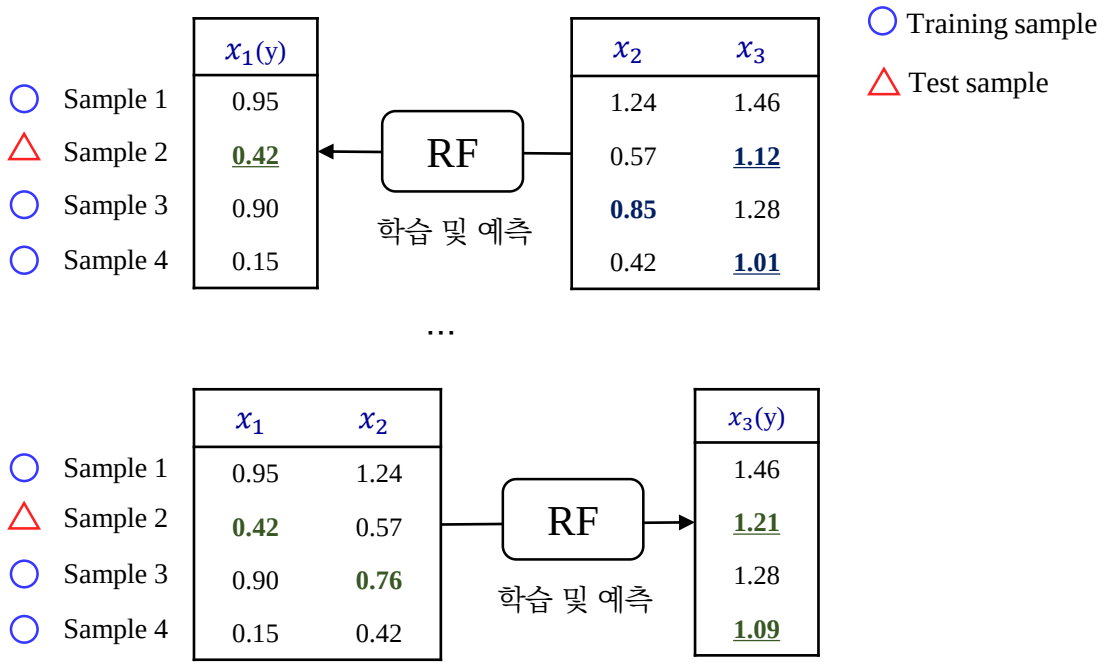


2 Related works

MissForest

- MissForest는 single imputation 기반의 결측치 대치 방법으로 random forest의 예측치를 결측치 대치 값으로 활용
- 총 4가지 step이 존재함
- Step 3) 직전 과정을 통해 결측이 보완된 데이터에 동일한 과정을 진행

MissForest 알고리즘 전개	
Step 1	평균이나 다른 대치 방법을 사용하여 결측치를 채움
Step 2	결측치 비율이 가장 낮은 변수부터 Random Forest를 사용하여 예측을 진행함. 이 때 학습 데이터는 결측이 발생하지 않은 샘플이며 이를 토대로 결측치에 대해 예측을 진행함
Step 3	모든 변수에 대해서 예측이 종료되면 Step2의 결측이 보완된 데이터를 기반으로 다시 변수 별 예측을 진행함
Step 4	추정된 결측치의 변화량이 특정 범위 이내로 수렴할 때까지 Step2와 Step3을 반복. 수렴 조건을 만족하면 이를 최종 데이터로써 활용함

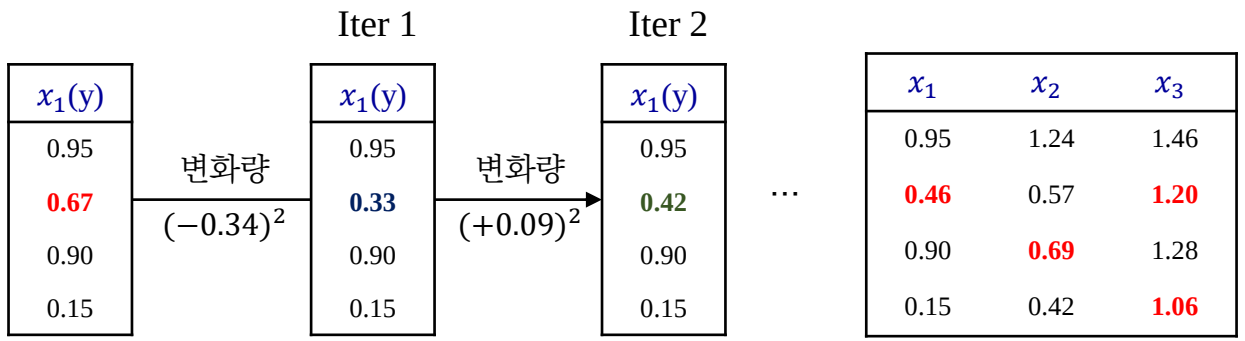


2 Related works

MissForest

- MissForest는 single imputation 기반의 결측치 대치 방법으로 random forest의 예측치를 결측치 대치 값으로 활용
- 총 4가지 step이 존재함
- Step 4) 결측 대체 값의 변화량이 수렴할 때까지 Step 2, 3을 반복

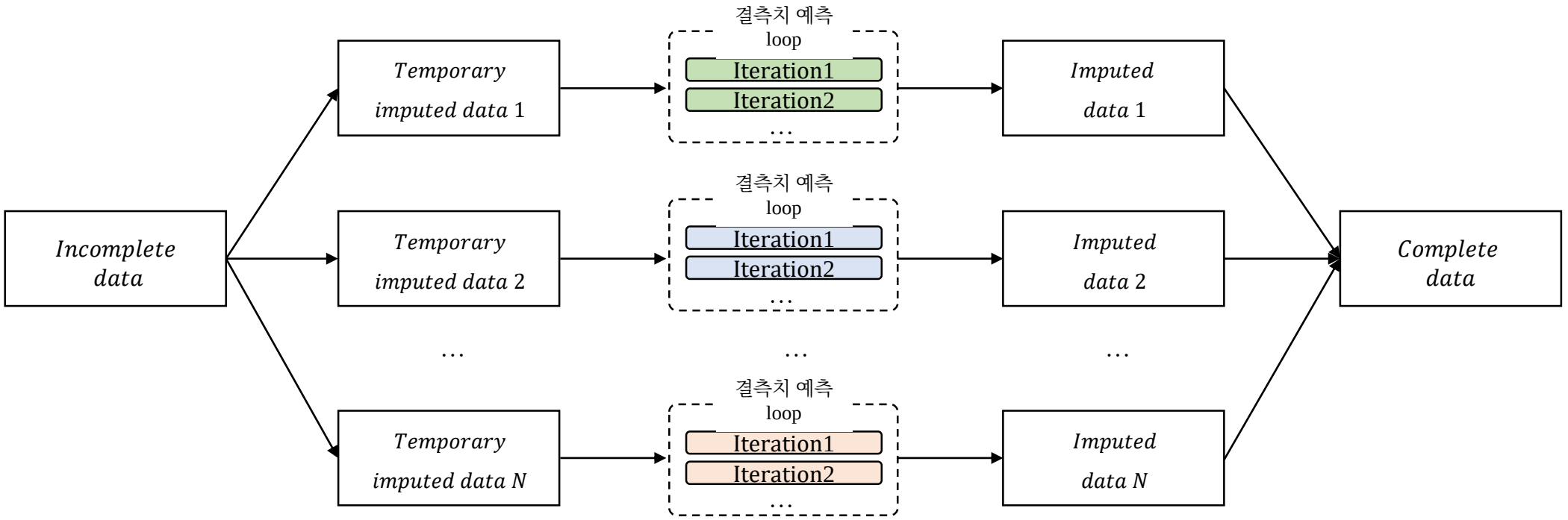
MissForest 알고리즘 전개	
Step 1	평균이나 다른 대치 방법을 사용하여 결측치를 채움
Step 2	결측치 비율이 가장 낮은 변수부터 Random Forest를 사용하여 예측을 진행함. 이 때 학습 데이터는 결측이 발생하지 않은 샘플이며 이를 토대로 결측치에 대해 예측을 진행함
Step 3	모든 변수에 대해서 예측이 종료되면 Step2의 결측이 보완된 데이터를 기반으로 다시 변수 별 예측을 진행함
Step 4	추정된 결측치의 변화량이 특정 범위 이내로 수렴할 때까지 Step2와 Step3을 반복. 수렴 조건을 만족하면 이를 최종 데이터로써 활용함



Related works

MICE : Multiple Imputation by Chained Equations

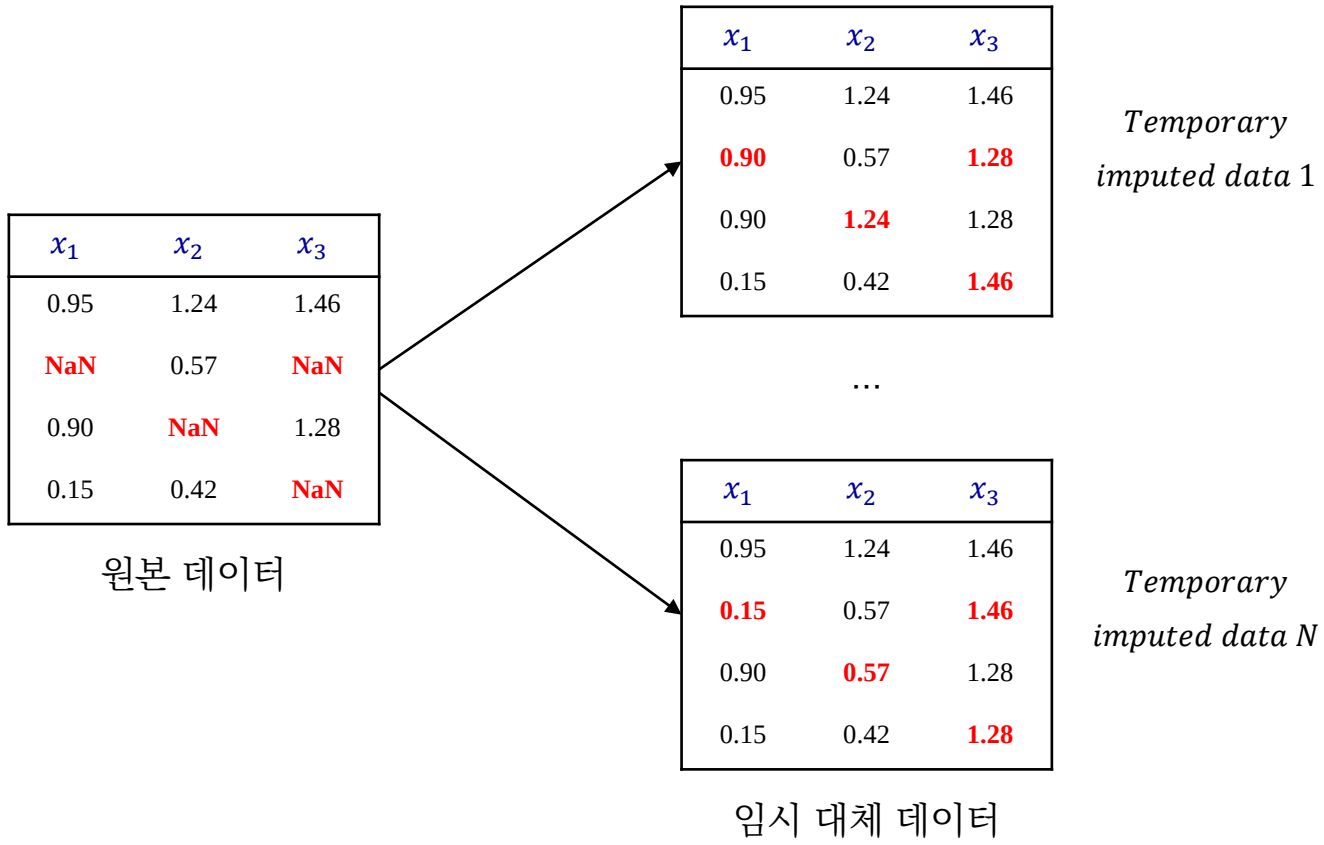
- MICE는 multiple imputation 방법 중 하나로 n개의 데이터에 MissForest와 유사한 과정을 거침



2 Related works

MICE : Multiple Imputation by Chained Equations

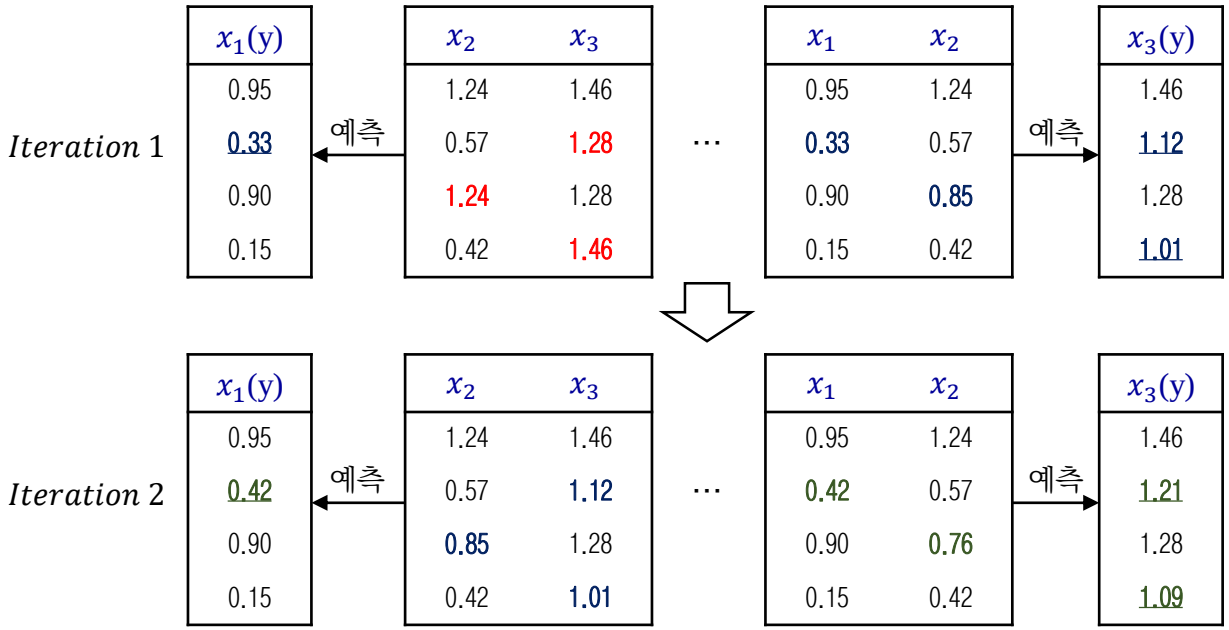
- 임시 대체의 경우 분포로부터의 샘플링 혹은 기존 관측치의 단순 복사를 통해서 임의로 결측치를 대체



Related works

MICE : Multiple Imputation by Chained Equations

- 변수 별 결측치 예측 단계에서는 임시로 채워진 데이터를 바탕으로 순차적인 예측을 통해 결측치를 보완 (MissForest와 동일)

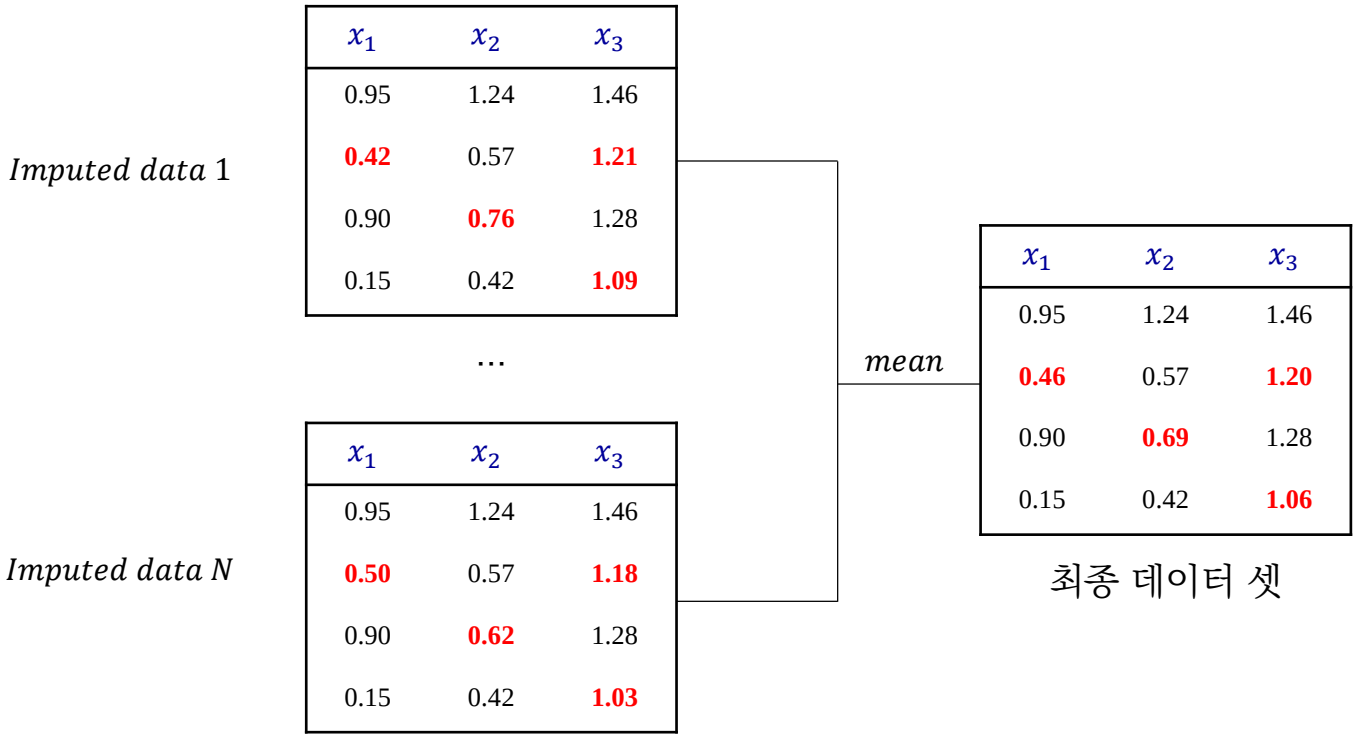


결측치 예측 기법	변수 형태
Linear regression	Numeric
Bayesian linear regression	Numeric
Logistic regression	Categorical
Linear discriminant analysis	Categorical
Predictive mean matching	Any
Random sample from observed values	Any
CART	Any
Random forest	Any

Related works

MICE : Multiple Imputation by Chained Equations

- 3) 종합 과정에서는 결측치가 모두 채워진 n개의 데이터의 평균을 내어 최종 데이터 셋을 구축
- 이 외에도 n개의 데이터 셋을 통해 표준 오차(standard error), 신뢰 구간(confidence interval) 등을 구할 수 있음



2 SIL Related works

MIDA : Multiple Imputation using Denoising Autoencoders

- GAN을 활용한 모델 이전, DAE를 활용한 multiple imputation method가 제안됨
- DAE는 오토 인코더의 파생 모델로 noise가 섞인 입력 데이터에서 이를 제거하여 원래의 입력 데이터로 복원하는 모델

x_1	x_2	x_3
0.95 (+0.03)	1.24 (-0.14)	1.46 (+0.27)
0.42 (-0.16)	0.57 (-0.21)	1.21 (+0.08)
0.90 (+0.05)	0.76 (+0.08)	1.28 (-0.11)
0.15 (-0.18)	0.42 (+0.31)	1.09 (-0.21)

잡음이 섞인 입력 데이터

encoder

z_1	z_2
0.89	-0.67
-1.62	0.88
0.05	-0.48
-0.24	1.72

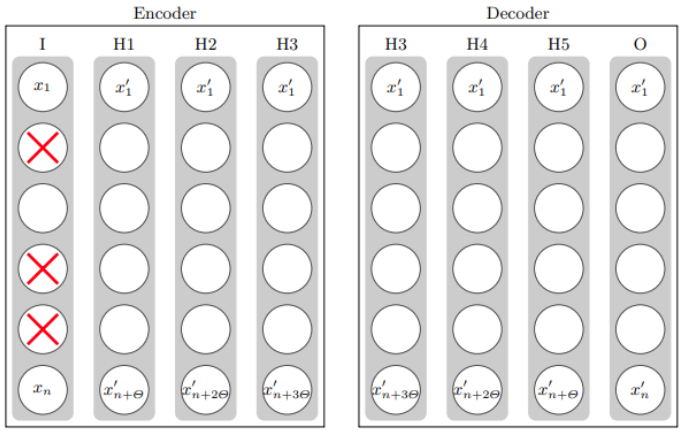
잠재 변수

decoder

x_1	x_2	x_3
0.95	1.24	1.46
0.42	0.57	1.21
0.90	0.76	1.28
0.15	0.42	1.09

복원 데이터

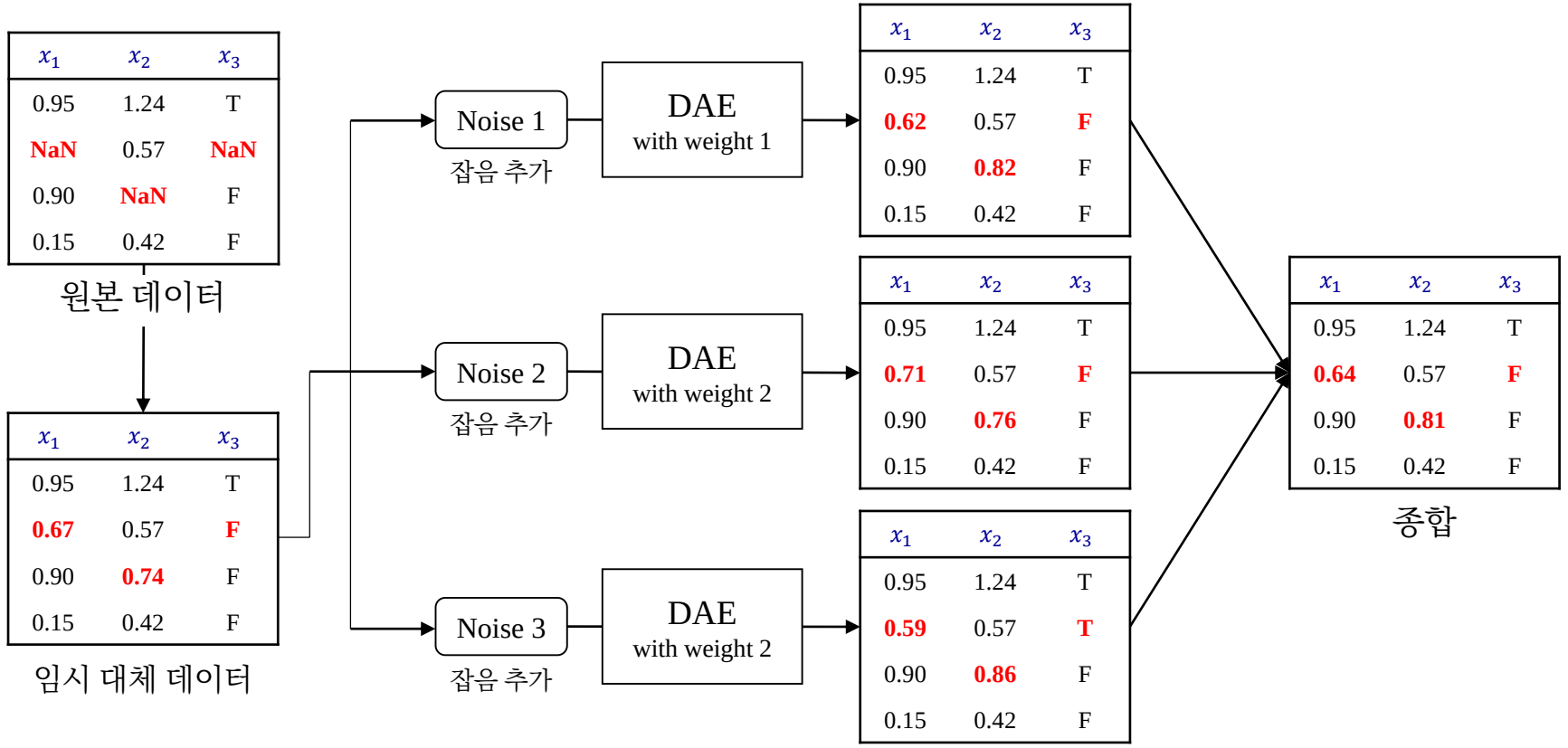
- MIDA는 잠재 변수의 차원이 입력 데이터의 차원보다 높은 overcomplete 오토 인코더 구조를 사용
  → 해당구조는 입력 데이터가 제공하는 정보보다 더 많은 정보를 찾을 때 사용함
- 레이어 당 θ만큼 노드가 추가되며 디코더는θ만큼 노드가 감소함 (θ기본값 = 7)



2 Related works

MIDA : Multiple Imputation using Denoising Autoencoders

- 먼저 수치형 변수인 경우 평균, 범주형 변수인 경우 최빈값을 사용하여 임시 대치를 진행
- 이후 DAE를 활용한 대치를 진행하며 데이터마다 DAE의 초기 가중치를 변경하여 다양한 결측 대체값을 도출 후 이를 종합



2 SIL Related works

MIDA : Multiple Imputation using Denoising Autoencoders

- UCI 데이터 셋 15개를 사용하였으며 다음과 같이 전체 데이터의 20%에 결측치를 발생시킴

	완전 무작위로 결측치 발생	중위수 이상 혹은 이하의 값에 결측치 발생
모든 변수에 적용	MCAR-Uniform	MNAR-Uniform
절반의 변수에 적용	MCAR-Random	MNAR-Random

- 아래의 표는 MICE와의 root mean square error (RMSE) 성능 비교 표 5개의 다중 대치 데이터의 평균(최소, 최대) RMSE를 의미함

	Data	Uniform missingness		Random missingness	
		DAE	MICE	DAE	MICE
MCAR	BH	2.9(2.9,3)	3.7(3.5,3.8)	0.9(0.9,1)	0.9(0.7,1)
	BC	2.9(2.9,2.9)	3.9(3.6,4.2)	1.2(1.2,1.3)	1.3(1.1,1.4)
	DN	25.7(25.7,25.7)	36.5(36.3,36.6)	13.1(13.1,13.2)	16.9(16.9,17)
	GL	1.1(1,1.1)	1.5(1.3,1.7)	1.3(1.2,1.4)	1.4(1.3,1.6)
	HV	2.4(2.4,2.4)	3.4(3.1,3.7)	1.1(1.1,1.2)	1.2(0.9,1.3)
	IS	13(12.9,13.1)	17.1(16.2,17.7)	5.8(5.6,6.2)	7(6.7,7.5)
	ON	2.1(2.1,2.1)	3.1(3,3.3)	0.9(0.9,1)	1(1,1.2)
	SL	3.6(3.6,3.7)	4.5(4.4,4.6)	1.8(1.7,1.8)	0.7(0.7,0.7)
	SR	1.2(1.,1.2)	1.5(1.4,1.7)	0.4(0.4,0.5)	0.4(0.4,0.5)
	ST	16.5(16.5,16.7)	27.9(27.5,28.2)	6.5(6.4,6.7)	13(12.5,13.8)
	SN	5.1(5,5.1)	7.3(7.2,7.5)	2.3(2.2,2.3)	3.2(3.2,3.3)
	SB	1.8(1.8,1.8)	2.4(2.3,2.4)	1.2(1.1,1.2)	1.1(1,1.1)
	VC	4.1(4,4.1)	5.6(5.5,5.7)	1.6(1.6,1.6)	2.2(2.1,2.3)
	VW	5.8(5.7,6.2)	7.7(7,8.1)	2.6(2.4,2.9)	3.8(3.3,4.2)
	ZO	2.1(2.1,2.1)	3.4(3.1,4.3)	1.1(1.1,1.2)	1.1(1.1,1.1)

	Data	Uniform missingness		Random missingness	
		DAE	MICE	DAE	MICE
MNAR	BH	2.3(2.2,2.4)	3.2(2.9,3.4)	0.9(0.8,1)	0.7(0.7,0.8)
	BC	2.9(2.8,3)	3.6(3.4,3.8)	1.7(1.7,1.8)	1.4(1.3,1.5)
	DN	25.3(25.2,25.3)	34.5(34.5,34.7)	5.7(5.7,5.8)	7.2(7.1,7.2)
	GL	1.3(1.3,1.4)	1.5(1.3,1.8)	0.4(0.3,0.4)	0.2(0.11,0.2)
	HV	2.6(2.6,2.6)	3.5(3.3,3.7)	1.3(1.2,1.3)	1.3(1.3,1.4)
	IS	11.7(11.5,11.8)	15.4(14.9,16.5)	4.8(4.5,5.1)	6.3(5.6,6.8)
	ON	1.5(1.5,1.5)	2.2(2,2.4)	1.2(1.1,1.2)	1.3(1.1,1.5)
	SL	3.4(3.4,3.4)	3.8(3.8,3.9)	1.6(1.6,1.6)	0.5(0.5,0.5)
	SR	1.2(1.2,1.2)	1.6(1.5,1.7)	0.4(0.3,0.4)	0.3(0.2,0.3)
	ST	11.8(11.7,11.9)	22.4(22.1,22.7)	4.5(4.3,4.7)	9.5(8.4,10.3)
	SN	4.6(4.6,4.6)	6.8(6.5,7.1)	2.3(2.3,2.4)	3.1(3,3.2)
	SB	1.7(1.7,1.7)	2.3(2.2,2.4)	0.6(0.6,0.6)	0.9(0.9,0.9)
	VC	3.5(3.4,3.7)	4.6(4.4,4.8)	1.7(1.7,1.8)	2.4(2.3,2.4)
	VW	5.9(5.9,5.9)	NA	2.3(2.1,2.5)	NA
	ZO	3.3(2.8,5.5)	3.9(3.6,4.6)	0.9(0.8,1.0)	1.1(0.7,1.7)

- MNAR-Random을 제외한 모든 상황에서 MIDA의 성능이 좋은 것을 확인할 수 있음
- 또한 MICE의 경우 특정 변수에서 결측치가임계치 이상으로 발생하면 예측을 하지 않기에 MNAR-VW에서 NA값이 발생한 반면 DAE는 그에 상관없이 결측치 보완이 가능하다는 장점 또한 존재함 (과연 이게 사실인가?)

3 SIL Proposed methodology

GAIN : Generative Adversarial Imputation Nets

- Generative Adversarial Imputation Nets (GAINs, 2018 ICML)는 GANs의 구조를 변경하여 결측치를 보완하는 방법임

✓ 생성자 부분

- 생성자는 3가지 input을 받음
- 결측치 생성이 완료된 행렬 (Imputed matrix)을 출력

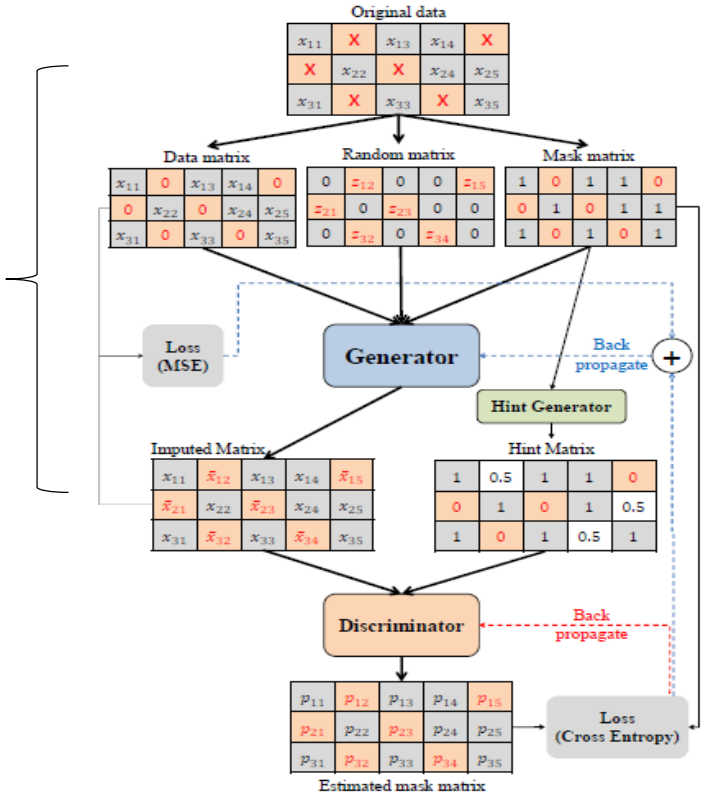


Figure 1. The architecture of GAIN

✓ 판별자 부분

- 판별자는 2가지 input을 받음
- Imputed matrix의 element별로 진짜인지 가짜인지 분류하는 역할

3 Proposed methodology

GAIN : Generator

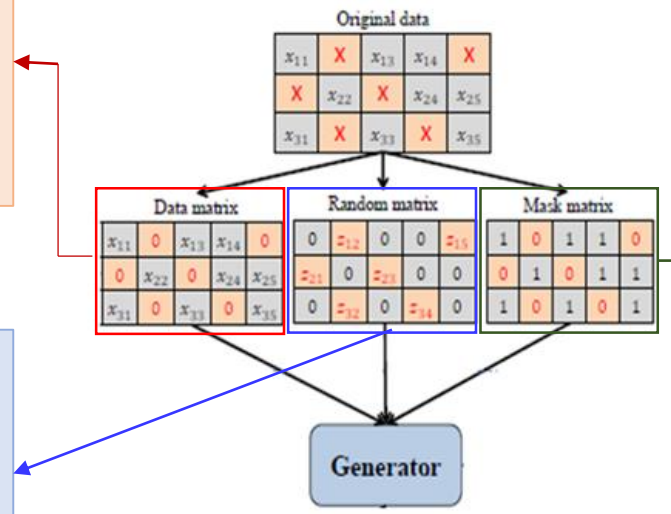
- 생성자의 3가지 input은 결측치가 존재하는 원본 데이터(original data)를 변형한 행렬이며 각각의 용도는 아래와 같음

✓ **Data matrix ($\tilde{X}$)**

- 기존값은 유지 후 결측치만0으로 대치한 행렬
- 생성자가 **만들고자하는 목표 분포에 해당**하며 정답 역할을 함

✓ **Random matrix ($(1-M)\odot Z$)**

- 결측치를 임의의 난수로 대치하며 기존값은0으로 대치한 행렬
- 생성자가 **이를 토대로 결측치를 생성** (GANs의 잠재변수와 동일)



✓ **Mask matrix (M)**

- 기존값은1로결측치는0으로 대치함
- 생성자와판별자 모두 사용됨
- 주된 목적은**판별자의 정답지** 및 판별자의 학습을 돕는 **hint생성**임
- 생성자에서의 역할은 $(1-M)\odot Z$을 통해 차원을 일치시키는 역할을 함 ($\odot$는 pairwise product)

Proposed methodology

GAIN : Generator

- 앞에서 정의한 3가지 input을 받아 생성자(G)는 결측치를 보완한 행렬 ($\bar{X}$)을 출력함

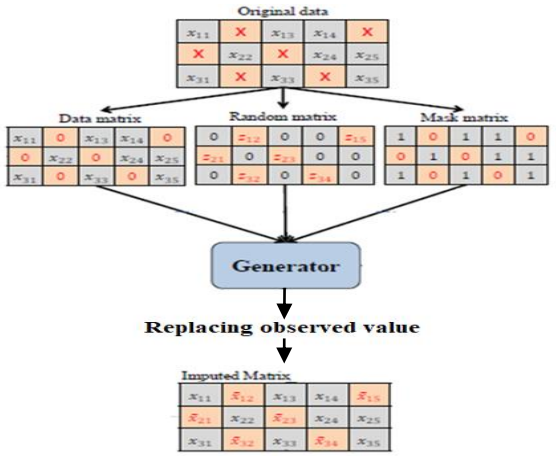
$$\bar{X} = G(\tilde{X}, M, (\mathbf{1} - M) \odot Z) \qquad (1)$$

- 그러나 생성자를 거치면서 신경망 연산에 의해 기존값($\tilde{x}_{ij}$) 또한 변형이 발생함
- 따라서 기존값은 그대로 사용하고 결측치만 생성값으로 대체한 Imputed matrix ($\hat{X}$)를 최종적으로 사용함

$$\hat{X} = M \odot \tilde{X} + (\mathbf{1} - M) \odot \bar{X} \qquad (2)$$

생성자의 inputs

결측치보완 행렬($\hat{X}$)
(=최종적으로 사용할 데이터)

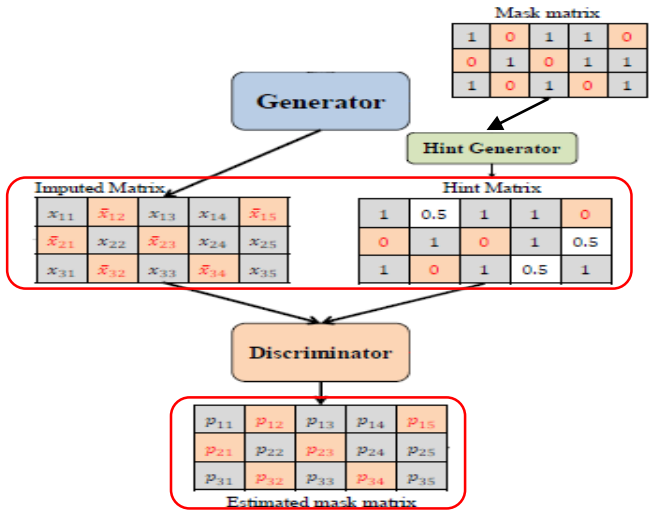


3 Proposed methodology

GAIN : Discriminator

- 판별자의 input은 두가지로 생성자를 거친 imputed matrix ($\hat{X}$)와 mask matrix를 토대로 생성된 hint matrix임
- Hint matrix는 판별자의 원활한 학습을 유도하기 위해 특정 element에 대한 정답을 제공하며 epoch마다 임의로 초기화 됨 (= Hint generator)
- 아래 그림에서 hint matrix 중 0과 1은 답이 제공되어 있으며 0.5로 할당된 element에 대해서 분류 학습을 진행함
- 판별자의 output은 생성된 imputed matrix의 개별 element에 대한 실제와 가짜의 예측 확률값임

판별자의 input



판별자의 output
(=element 별 예측 확률)

3 Proposed methodology

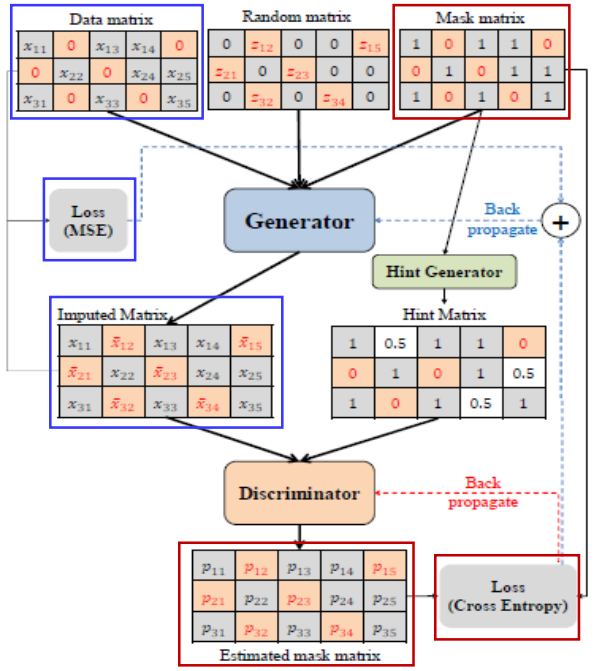
GAIN : Loss

- 생성자의 목적함수는 두가지 항으로 구성되며 기존값($\tilde{\boldsymbol{x}}_{ij}$)과 생성자를 거쳐 변형이 발생한 값 ($\bar{\boldsymbol{x}}_{ij}$)의 MSE가 첫번째 loss 항이 됨

$$\mathcal{L}_{G1} = MSE(\tilde{X}, M \odot \bar{X}) \qquad (3)$$

- 판별자의output인 element별 예측 확률과 정답인 mask matrix 사이의 cross entropy는 판별자의loss임과 동시에 생성자의 두번째 loss로 적용됨

✓ 생성자의 첫번째 loss



✓ 판별자의 loss

✓ 생성자의 두번째 loss

Results

실험 결과

- 5가지 데이터 셋에 실험한 결과 RMSE 척도를 기준으로 MIDA (Auto-encoder)를 포함한 기존의 방법론과 비교하여 우수함

Table 2. Imputation performance in terms of RMSE (Average ± Std of RMSE)

Algorithm	Breast	Spam	Letter	Credit	News
GAIN	**.0546 ± .0006**	**.0513± .0016**	**.1198± .0005**	**.1858 ± .0010**	**.1441 ± .0007**
MICE	.0646 ± .0028	.0699 ± .0010	.1537 ± .0006	.2585 ± .0011	.1763 ± .0007
MissForest	.0608 ± .0013	.0553 ± .0013	.1605 ± .0004	.1976 ± .0015	.1623 ± 0.012
Matrix	.0946 ± .0020	.0542 ± .0006	.1442 ± .0006	.2602 ± .0073	.2282 ± .0005
Auto-encoder	.0697 ± .0018	.0670 ± .0030	.1351 ± .0009	.2388 ± .0005	.1667 ± .0014
EM	.0634 ± .0021	.0712 ± .0012	.1563 ± .0012	.2604 ± .0015	.1912 ± .0011

- 데이터의 결측률에 따른 실험(a)과 변수의 개수에 따른 실험(c)에서 기존 방법 대비 좋은 성능을 보임
- 다만 데이터 개수에 따른 실험(b)에서 여타의 신경망과 비슷하게 충분한 데이터가 확보되어야 좋은 성능을 얻을 수 있다는 한계점도 존재함

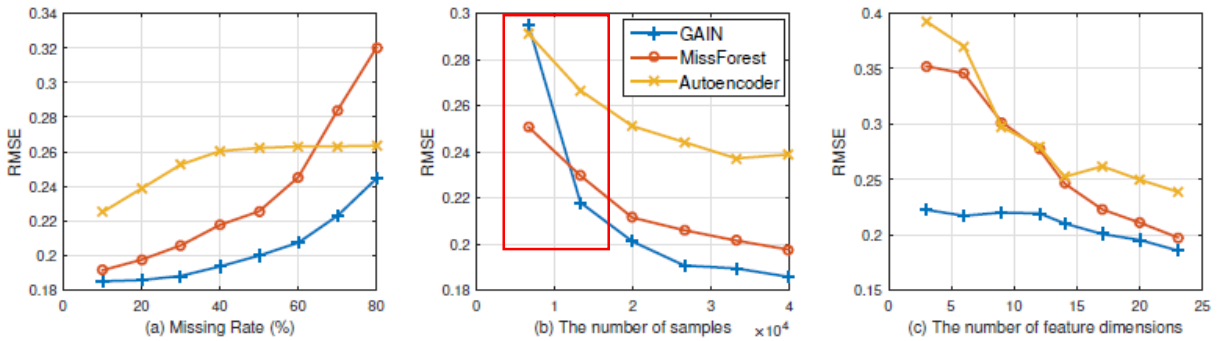


Figure 2. RMSE performance in different settings: (a) Various missing rates, (b) Various number of samples, (c) Various feature dimensions

Q&A