UNSUPERVISED DOMAIN ADAPTATION WITH NON-STOCHASTIC MISSING DATA

ECML 2021 - Data Mining and Knowledge Discovery

Monday 13th September, 2021 to Friday 17th September, 2021

Matthieu Kirchmeyer[1,2], Patrick Gallinari[1,2], Alain Rakotomamonjy[2,3], Amin Mantrach[4]

[1]Sorbonne Université, CNRS, LIP6, [2]Criteo AI Lab,
[3]Université de Rouen - LITIS, [4]Amazon

Introduction

Missing data

SCIENCES SORBONNE UNIVERSITÉ

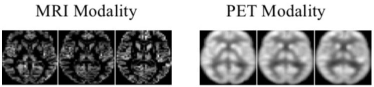
- Missing data is present in many real-world applications.

Missing data

- Missing data is present in many real-world applications.

Missing data

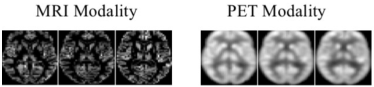
- Missing data is present in many real-world applications.

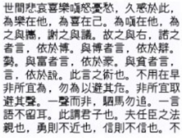


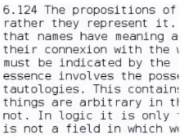
Missing data

SCIENCES SORBONNE UNIVERSITÉ

- Missing data is present in many real-world applications.

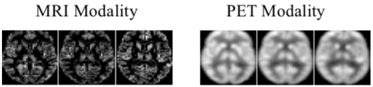


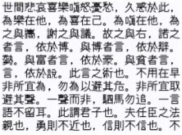


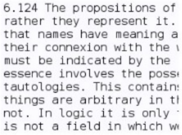


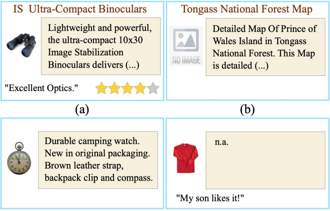
Missing data

- Missing data is present in many real-world applications.





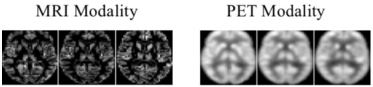


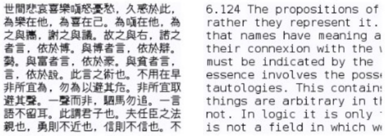


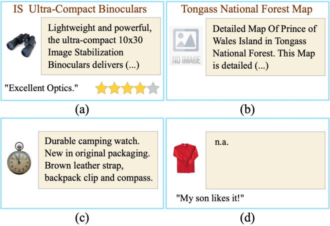
(a) (b)
(c) (d)

Missing data

- Missing data is present in many real-world applications.



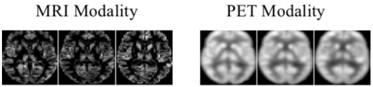


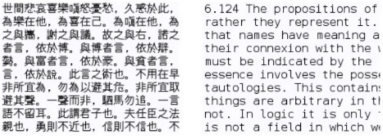


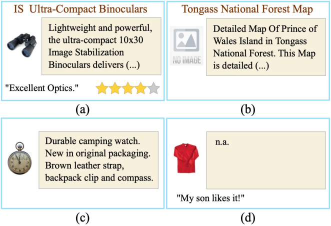
- Existing methods usually consider **stochastic missing data**.

Missing data

- Missing data is present in many real-world applications.







- Existing methods usually consider **stochastic missing data**.
- Missing Completely At Random (MCAR) Rubin 1976

$$\forall x,\ p_\phi(m|x) = p_\phi(m)$$

m **stochastic**.

Non-stochastic missing data

- MCAR when m is **deterministic**, a.k.a. **non-stochastic missing data**, is seldom considered.

Non-stochastic missing data

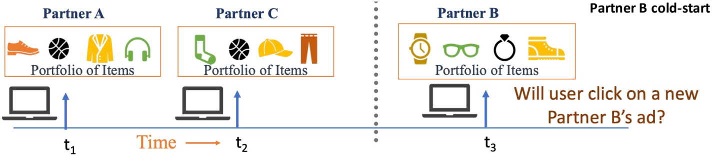
SCIENCES SORBONNE UNIVERSITÉ

- MCAR when m is **deterministic**, a.k.a. **non-stochastic missing data**, is seldom considered.
- Yet, common in applications e.g. cold-start

Non-stochastic missing data

SCIENCES SORBONNE UNIVERSITÉ

- MCAR when m is **deterministic**, a.k.a. **non-stochastic missing data**, is seldom considered.
- Yet, common in applications e.g. cold-start



Non-stochastic missing data

SCIENCES SORBONNE UNIVERSITÉ

- MCAR when m is **deterministic**, a.k.a. **non-stochastic missing data**, is seldom considered.
- Yet, common in applications e.g. cold-start

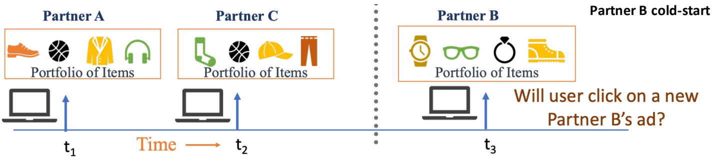
Contributions

- Handle non-stochastic missing data with unsupervised domain adaptation (UDA).

Non-stochastic missing data

SCIENCES SORBONNE UNIVERSITÉ

- MCAR when m is **deterministic**, a.k.a. **non-stochastic missing data**, is seldom considered.
- Yet, common in applications e.g. cold-start



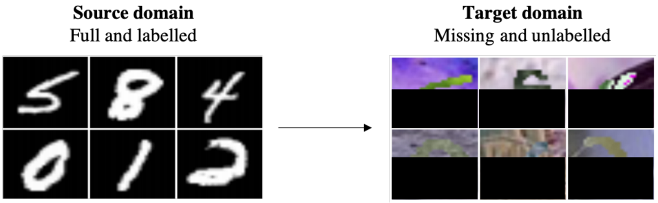
Contributions

- Handle non-stochastic missing data with unsupervised domain adaptation (UDA).
- Formalize the problem.

Adaptation-Imputation problem definition

SCIENCES SORBONNE UNIVERSITÉ

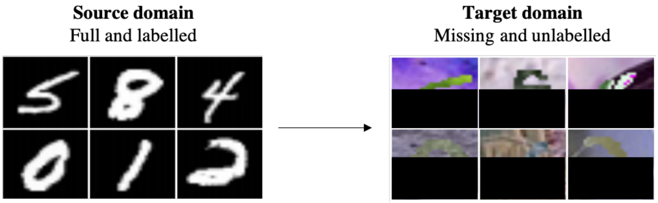
1. labelled x_S and unlabelled x_T under distribution shift.



Adaptation-Imputation problem definition

SCIENCES SORBONNE UNIVERSITÉ

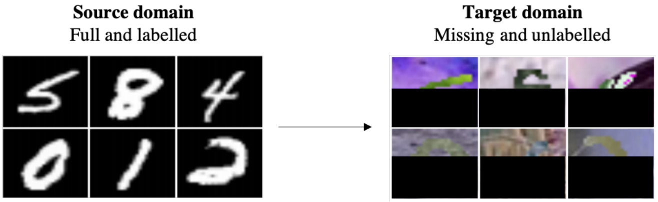
1. labelled x_S and unlabelled x_T under distribution shift.
2. $x_e = (x_{e_1}, x_{e_2}), e \in \{S, T\}$ with x_S fully observed; x_{T_2} **missing**.



Adaptation-Imputation problem definition

1. labelled x_S and unlabelled x_T under distribution shift.
2. $x_e = (x_{e_1}, x_{e_2}), e \in \{S, T\}$ with x_S fully observed; x_{T_2} **missing**.

(1), (2) $\rightarrow$ **UDA** under **non-stochastic** missingness.

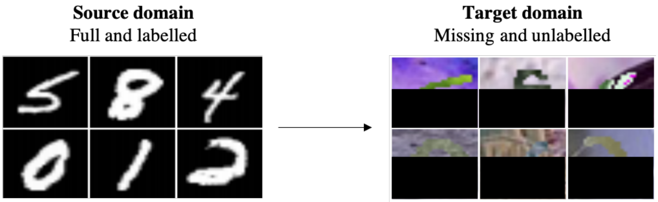


Adaptation-Imputation problem definition

1. labelled x_S and unlabelled x_T under distribution shift.
2. $x_e = (x_{e_1}, x_{e_2}), e \in \{S, T\}$ with x_S fully observed; x_{T_2} **missing**.
3. no supervision for imputation on T.

(1), (2) $\rightarrow$ **UDA** under **non-stochastic** missingness.

(3) $\rightarrow$ **imputation** without supervision.

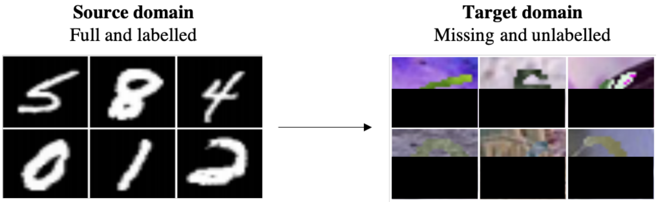


Adaptation-Imputation problem definition

1. labelled x_S and unlabelled x_T under distribution shift.
2. $x_e = (x_{e_1}, x_{e_2}), e \in \{S, T\}$ with x_S fully observed; x_{T_2} **missing**.
3. no supervision for imputation on T.

(1), (2) $\rightarrow$ **UDA** under **non-stochastic** missingness.

(3) $\rightarrow$ **imputation** without supervision.



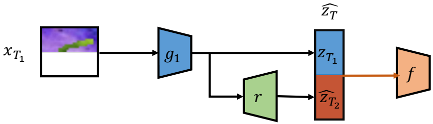
Goal: train a classifier $\hat{h}$ with low classification error on T.

Model

Model components

SCIENCES SORBONNE UNIVERSITÉ

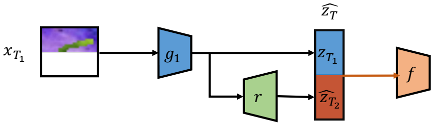
Model: $\hat{h} : \mathcal{X}_1 \to \mathcal{Y} = \{0, \ldots, K\}$, $\hat{h} = f \circ \hat{g}$ **on S and T**



Model components

Model: $\hat{h} : \mathcal{X}_1 \to \mathcal{Y} = \{0, \ldots, K\}$, $\hat{h} = f \circ \hat{g}$ **on S and T**

- $\hat{g} : \mathcal{X}_1 \to \mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$ encoder using $\mathrm{x}_{\mathbf{e_1}}$.

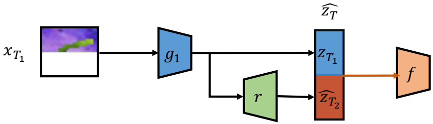


Model components

SCIENCES SORBONNE UNIVERSITÉ

Model: $\hat{h} : \mathcal{X}_1 \rightarrow \mathcal{Y} = \{0, \ldots, K\}$, $\hat{h} = f \circ \hat{g}$ **on S and T**

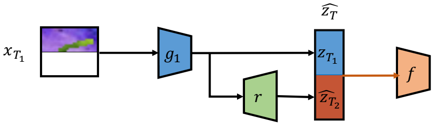
- $\hat{g} : \mathcal{X}_1 \rightarrow \mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$ encoder using $\mathrm{x}_{\mathbf{e_1}}$.
    - $g_1 : \mathcal{X}_1 \rightarrow \mathcal{Z}_1$ encoder of $\mathrm{x}_{\mathbf{e_1}}$.



Model components

Model: $\hat{h} : \mathcal{X}_1 \to \mathcal{Y} = \{0, \ldots, K\}$, $\hat{h} = f \circ \hat{g}$ **on S and T**

- $\hat{g} : \mathcal{X}_1 \to \mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$ encoder using $\mathrm{x}_{\mathbf{e_1}}$.
  - $g_1 : \mathcal{X}_1 \to \mathcal{Z}_1$ encoder of $\mathrm{x}_{\mathbf{e_1}}$.
  - $r : \mathcal{Z}_1 \to \mathcal{Z}_2$ conditional generator of $\mathrm{z}_{\mathbf{e_2}}$ given $\mathrm{z}_{\mathbf{e_1}}$.

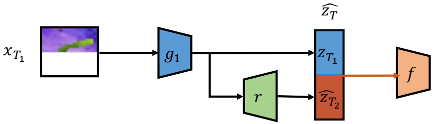


Model components

SCIENCES SORBONNE UNIVERSITÉ

Model: $\hat{h} : \mathcal{X}_1 \to \mathcal{Y} = \{0, \dots, K\}$, $\hat{h} = f \circ \hat{g}$ **on S and T**

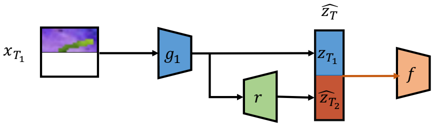
- $\hat{g} : \mathcal{X}_1 \to \mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$ encoder using $\mathrm{x}_{\mathbf{e_1}}$.
  - $g_1 : \mathcal{X}_1 \to \mathcal{Z}_1$ encoder of $\mathrm{x}_{\mathbf{e_1}}$.
  - $r : \mathcal{Z}_1 \to \mathcal{Z}_2$ conditional generator of $\mathrm{z}_{\mathbf{e_2}}$ given $\mathrm{z}_{\mathbf{e_1}}$.
- $f : \mathcal{Z} \to \mathcal{Y} = \{0, \dots, K\}$ classifier.



Model components

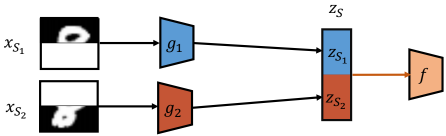
Model: $\hat{h} : \mathcal{X}_1 \to \mathcal{Y} = \{0, \dots, K\}$, $\hat{h} = f \circ \hat{g}$ **on S and T**

- $\hat{g} : \mathcal{X}_1 \to \mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$ encoder using $\mathrm{x}_{\mathbf{e_1}}$.
  - $g_1 : \mathcal{X}_1 \to \mathcal{Z}_1$ encoder of $\mathrm{x}_{\mathbf{e_1}}$.
  - $r : \mathcal{Z}_1 \to \mathcal{Z}_2$ conditional generator of $\mathrm{z}_{\mathbf{e_2}}$ given $\mathrm{z}_{\mathbf{e_1}}$.
- $f : \mathcal{Z} \to \mathcal{Y} = \{0, \dots, K\}$ classifier.



Reference: $h : \mathcal{X} \to \mathcal{Y}$, $h = f \circ g$ **only on S**

- $g : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Z}$ encoder with both $(\mathrm{x}_{\mathbf{e_1}}, \mathrm{x}_{\mathbf{e_2}})$ components.
  - $g_2 : \mathcal{X}_2 \to \mathcal{Z}_2$ encoder of $\mathrm{x}_{\mathbf{e_2}}$.



Model training

Adversarial training

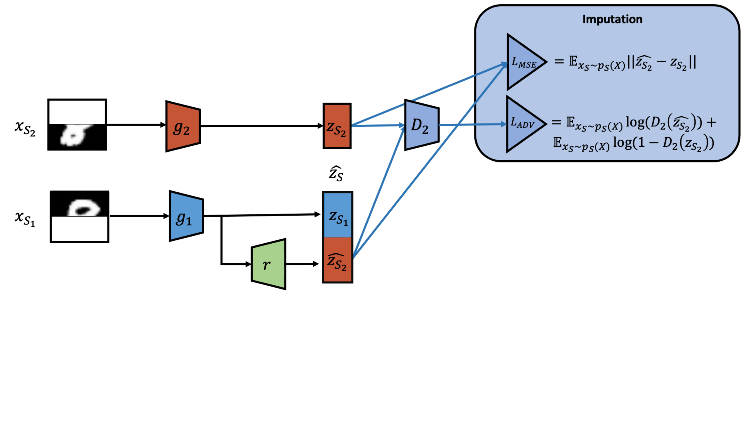
Three modules for imputation, adaptation, classification.

Model training

SCIENCES SORBONNE UNIVERSITÉ

Adversarial training

- **Imputation**: alignment loss $L_2 = L_{ADV} + \lambda_{MSE} L_{MSE}$.

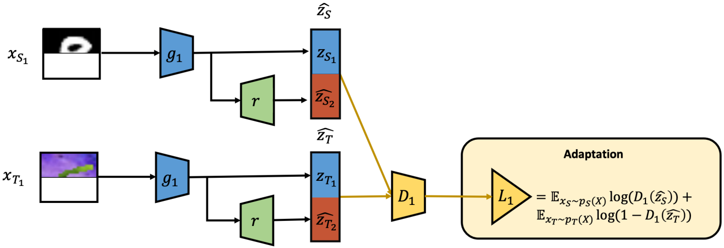


Model training

SCIENCES SORBONNE UNIVERSITÉ

Adversarial training

- **Adaptation**: alignment loss L_1.

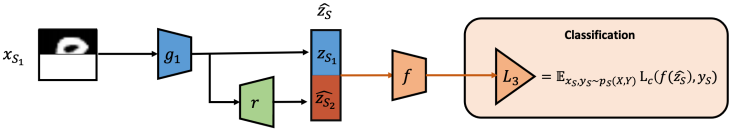


Model training

SCIENCES SORBONNE UNIVERSITÉ

Adversarial training

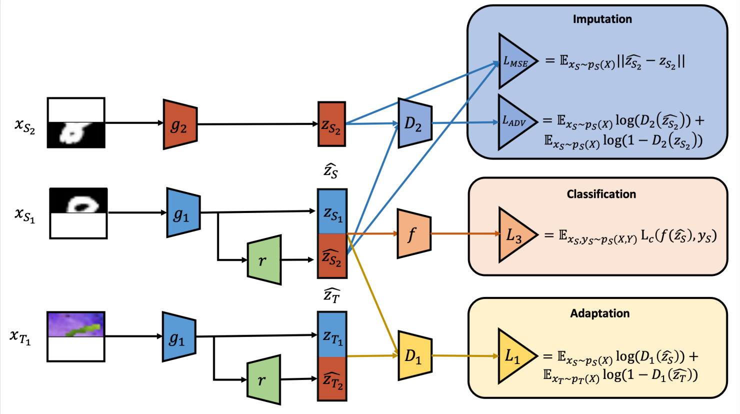
- **Classification**: cross-entropy loss L_3.



Model training

SCIENCES SORBONNE UNIVERSITÉ

$$\min_{g_1, g_2, r, f} \max_{D_1, D_2} L_1 + (L_{ADV} + \lambda_{MSE} L_{MSE}) + L_3 \tag{1}$$



Formalization

Assumptions

Conditional invariance

After projection with $g = (g_1, g_2)$,

$$p_S(Z_2|Z_1) = p_T(Z_2|Z_1), \quad p_S(Z_1) \neq p_T(Z_1)$$

Assumptions

SCIENCES SORBONNE UNIVERSITÉ

Conditional invariance

After projection with $g = (g_1, g_2)$,

$$p_S(Z_2|Z_1) = p_T(Z_2|Z_1), \quad p_S(Z_1) \neq p_T(Z_1)$$

We can use available supervision in S to infer $p_T(Z_2|Z_1)$.

Assumptions

SCIENCES SORBONNE UNIVERSITÉ

Conditional invariance

After projection with $g = (g_1, g_2)$,

$$p_S(Z_2|Z_1) = p_T(Z_2|Z_1), \quad p_S(Z_1) \neq p_T(Z_1)$$

We can use available supervision in S to infer $p_T(Z_2|Z_1)$.

Covariate shift

After projection with $\hat{g} = (g_1, r \circ g_1)$,

$$p_S(Y|\hat{Z}) = p_T(Y|\hat{Z}), \quad p_S(\hat{Z}) \neq p_T(\hat{Z})$$

Assumptions

Conditional invariance

After projection with $g = (g_1, g_2)$,

$$p_S(Z_2|Z_1) = p_T(Z_2|Z_1), \quad p_S(Z_1) \neq p_T(Z_1)$$

We can use available supervision in S to infer $p_T(Z_2|Z_1)$.

Covariate shift

After projection with $\hat{g} = (g_1, r \circ g_1)$,

$$p_S(Y|\hat{Z}) = p_T(Y|\hat{Z}), \quad p_S(\hat{Z}) \neq p_T(\hat{Z})$$

We can find $\hat{h} = f \circ \hat{g}$ with low source and target error; common assumption for UDA.

Assumptions

SCIENCES SORBONNE UNIVERSITÉ

Conditional invariance

After projection with $g = (g_1, g_2)$,

$$p_S(Z_2|Z_1) = p_T(Z_2|Z_1), \quad p_S(Z_1) \neq p_T(Z_1)$$

We can use available supervision in S to infer $p_T(Z_2|Z_1)$.

Covariate shift

After projection with $\hat{g} = (g_1, r \circ g_1)$,

$$p_S(Y|\hat{Z}) = p_T(Y|\hat{Z}), \quad p_S(\hat{Z}) \neq p_T(\hat{Z})$$

We can find $\hat{h} = f \circ \hat{g}$ with low source and target error; common assumption for UDA.

Upper-bounds

Assumptions

SCIENCES SORBONNE UNIVERSITÉ

Conditional invariance

After projection with $g = (g_1, g_2)$,

$$p_S(Z_2|Z_1) = p_T(Z_2|Z_1), \quad p_S(Z_1) \neq p_T(Z_1)$$

We can use available supervision in S to infer $p_T(Z_2|Z_1)$.

Covariate shift

After projection with $\hat{g} = (g_1, r \circ g_1)$,

$$p_S(Y|\hat{Z}) = p_T(Y|\hat{Z}), \quad p_S(\hat{Z}) \neq p_T(\hat{Z})$$

We can find $\hat{h} = f \circ \hat{g}$ with low source and target error; common assumption for UDA.

Upper-bounds

- Adaptation upper-bound of the target error of $\hat{h}$

Assumptions

SCIENCES SORBONNE UNIVERSITÉ

Conditional invariance

After projection with $g = (g_1, g_2)$,

$$p_S(Z_2|Z_1) = p_T(Z_2|Z_1), \quad p_S(Z_1) \neq p_T(Z_1)$$

We can use available supervision in S to infer $p_T(Z_2|Z_1)$.

Covariate shift

After projection with $\hat{g} = (g_1, r \circ g_1)$,

$$p_S(Y|\hat{Z}) = p_T(Y|\hat{Z}), \quad p_S(\hat{Z}) \neq p_T(\hat{Z})$$

We can find $\hat{h} = f \circ \hat{g}$ with low source and target error; common assumption for UDA.

Upper-bounds

- Adaptation upper-bound of the target error of $\hat{h}$
- Imputation upper-bound of the target error of h

Upper-bounds

SCIENCES SORBONNE UNIVERSITÉ

Adaptation upper-bound Ben-David et al. 2010

Given $f \in \mathcal{F}$ and $\hat{g}$

$$\epsilon_T(f \circ \hat{g}) \leq \underbrace{\left[\epsilon_S(f \circ \hat{g}) + d_{\mathcal{F}\Delta\mathcal{F}}(p_S(\hat{Z}), p_T(\hat{Z})) + \lambda_{\mathcal{H}_{\hat{g}}}\right]}_{\text{Domain Adaptation (DA)}} \quad (2)$$

Upper-bounds

Adaptation upper-bound Ben-David et al. 2010

Given $f \in \mathcal{F}$ and $\hat{g}$

$$\epsilon_T(f \circ \hat{g}) \leq \underbrace{\left[\epsilon_S(f \circ \hat{g}) + d_{\mathcal{F}\Delta\mathcal{F}}(p_S(\hat{Z}), p_T(\hat{Z})) + \lambda_{\mathcal{H}_{\hat{g}}}\right]}_{\text{Domain Adaptation (DA)}} \tag{2}$$

- $\epsilon_e(\cdot)$: expected error on $e \in \{S, T\}$

Upper-bounds

SCIENCES SORBONNE UNIVERSITÉ

Adaptation upper-bound Ben-David et al. 2010

Given $f \in \mathcal{F}$ and $\hat{g}$

$$\epsilon_T(f \circ \hat{g}) \leq \underbrace{\left[\epsilon_S(f \circ \hat{g}) + d_{\mathcal{F}\Delta\mathcal{F}}(p_S(\hat{Z}), p_T(\hat{Z})) + \lambda_{\mathcal{H}_{\hat{g}}}\right]}_{\text{Domain Adaptation (DA)}} \tag{2}$$

- $\epsilon_e(\cdot)$: expected error on $e \in \{S, T\}$
- $d_{\mathcal{F}\Delta\mathcal{F}}$: $\mathcal{F}\Delta\mathcal{F}$-divergence; $\mathcal{F}\Delta\mathcal{F}$: symmetric difference hypothesis space $h \in \mathcal{F}\Delta\mathcal{F} \iff \exists f_1, f_2 \in \mathcal{F}, h(x) = f_1(x) \oplus f_2(x)$

Upper-bounds

SCIENCES SORBONNE UNIVERSITÉ

Adaptation upper-bound Ben-David et al. 2010

Given $f \in \mathcal{F}$ and $\hat{g}$

$$\epsilon_T(f \circ \hat{g}) \leq \underbrace{\left[\epsilon_S(f \circ \hat{g}) + d_{\mathcal{F}\Delta\mathcal{F}}(p_S(\hat{Z}), p_T(\hat{Z})) + \lambda_{\mathcal{H}_{\hat{g}}}\right]}_{\text{Domain Adaptation (DA)}} \tag{2}$$

- $\epsilon_e(\cdot)$: expected error on $e \in \{S, T\}$
- $d_{\mathcal{F}\Delta\mathcal{F}}$: $\mathcal{F}\Delta\mathcal{F}$-divergence; $\mathcal{F}\Delta\mathcal{F}$: symmetric difference hypothesis space $h \in \mathcal{F}\Delta\mathcal{F} \iff \exists f_1, f_2 \in \mathcal{F}, h(x) = f_1(x) \oplus f_2(x)$
- $\lambda_{\mathcal{H}_{\hat{g}}}$: joint risk of the optimal hypothesis

$$\lambda_{\mathcal{H}_{\hat{g}}} = \min_{f' \in \mathcal{F}} \left[\epsilon_S(f' \circ \hat{g}) + \epsilon_T(f' \circ \hat{g})\right]$$

Upper-bounds

SCIENCES SORBONNE UNIVERSITÉ

Adaptation upper-bound Ben-David et al. 2010

Given $f \in \mathcal{F}$ and $\hat{g}$

$$\epsilon_T(f \circ \hat{g}) \leq \underbrace{\left[\epsilon_S(f \circ \hat{g}) + d_{\mathcal{F}\Delta\mathcal{F}}(p_S(\hat{Z}), p_T(\hat{Z})) + \lambda_{\mathcal{H}_{\hat{g}}}\right]}_{\text{Domain Adaptation (DA)}} \tag{2}$$

- $\epsilon_e(\cdot)$: expected error on $e \in \{S, T\}$
- $d_{\mathcal{F}\Delta\mathcal{F}}$: $\mathcal{F}\Delta\mathcal{F}$-divergence; $\mathcal{F}\Delta\mathcal{F}$: symmetric difference hypothesis space $h \in \mathcal{F}\Delta\mathcal{F} \iff \exists f_1, f_2 \in \mathcal{F}, h(x) = f_1(x) \oplus f_2(x)$
- $\lambda_{\mathcal{H}_{\hat{g}}}$: joint risk of the optimal hypothesis

$$\lambda_{\mathcal{H}_{\hat{g}}} = \min_{f' \in \mathcal{F}} \left[\epsilon_S(f' \circ \hat{g}) + \epsilon_T(f' \circ \hat{g})\right]$$

$L_3 \rightarrow$ 1st term, $L_1 \rightarrow$ 2nd term, Covariate Shift $\rightarrow$ 3rd term small.

Upper-bounds

Imputation upper-bound

Under Conditional Invariance, given $f, \hat{g}$ and g,

$$\epsilon_T(f \circ g) \leq \underbrace{\underbrace{\sup_{z \sim p(Z)} \left[\frac{p_S(Z_2 = z_2|z_1)}{p_S(\hat{Z}_2 = z_2|z_1)}\right]}_{\text{Imputation error on S } (I_S)} \times \underbrace{\sup_{z \sim p(Z)} \left[\frac{p_S(\hat{Z}_2 = z_2|z_1)}{p_T(\hat{Z}_2 = z_2|z_1)}\right]}_{\text{Transfer error of Imputation } (T_I)}}_{\text{Imputation error on T } (I_T)} \times \epsilon_T(f \circ \hat{g}) \tag{3}$$

Upper-bounds

SCIENCES SORBONNE UNIVERSITÉ

Imputation upper-bound

Under Conditional Invariance, given $f, \hat{g}$ and g,

$$\epsilon_T(f \circ g) \leq \underbrace{\underbrace{\sup_{z \sim p(Z)} [\frac{p_S(Z_2 = z_2|z_1)}{p_S(\hat{Z}_2 = z_2|z_1)}]}_{\text{Imputation error on S } (I_S)} \times \underbrace{\sup_{z \sim p(Z)} [\frac{p_S(\hat{Z}_2 = z_2|z_1)}{p_T(\hat{Z}_2 = z_2|z_1)}]}_{\text{Transfer error of Imputation } (T_I)}}_{\text{Imputation error on T } (I_T)} \times \epsilon_T(f \circ \hat{g}) \quad (3)$$

$L_2 \rightarrow (I_S)$, $L_1 \rightarrow (T_I)$, $(DA) \rightarrow$ 3rd term.

Experiments

Experimental setting

SCIENCES SORBONNE UNIVERSITÉ

Baselines

Experimental setting

Baselines

- `Full`: full x_S and x_T.

Experimental setting

Baselines

- `Full`: full x_S and x_T.
- `ZeroImputation`: full x_S; missing x_{T_2} set to 0, $x_T = (x_{T_1}, \mathbf{0})$.

Experimental setting

Baselines

- `Full`: full x_S and x_T.
- `ZeroImputation`: full x_S; missing x_{T_2} set to 0, $x_T = (x_{T_1}, \mathbf{0})$.
- `IgnoreComponent`: only x_{S_1}, x_{T_1}; x_{S_2}, x_{T_2} ignored.

Experimental setting

Baselines

- `Full`: full x_S and x_T.
- `ZeroImputation`: full x_S; missing x_{T_2} set to 0, $x_T = (x_{T_1}, \mathbf{0})$.
- `IgnoreComponent`: only x_{S_1}, x_{T_1}; x_{S_2}, x_{T_2} ignored.
- `Imputation`: full x_S; missing x_{T_2} imputed.

Experimental setting

SCIENCES SORBONNE UNIVERSITÉ

Baselines

- `Full`: full x_S and x_T.
- `ZeroImputation`: full x_S; missing x_{T_2} set to 0, $x_T = (x_{T_1}, \mathbf{0})$.
- `IgnoreComponent`: only x_{S_1}, x_{T_1}; x_{S_2}, x_{T_2} ignored.
- `Imputation`: full x_S; missing x_{T_2} imputed.
- Two divergences for aligning distributions:
  - $\mathcal{H}$-divergence
  - Wasserstein distance

Experimental setting

SCIENCES SORBONNE UNIVERSITÉ

Baselines

- `Full`: full x_S and x_T.
- `ZeroImputation`: full x_S; missing x_{T_2} set to 0, $x_T = (x_{T_1}, \mathbf{0})$.
- `IgnoreComponent`: only x_{S_1}, x_{T_1}; x_{S_2}, x_{T_2} ignored.
- `Imputation`: full x_S; missing x_{T_2} imputed.
- Two divergences for aligning distributions:
  - $\mathcal{H}$-divergence
  - Wasserstein distance

Datasets and Metrics

Experimental setting

SCIENCES SORBONNE UNIVERSITÉ

Baselines

- `Full`: full x_S and x_T.
- `ZeroImputation`: full x_S; missing x_{T_2} set to 0, $x_T = (x_{T_1}, \mathbf{0})$.
- `IgnoreComponent`: only x_{S_1}, x_{T_1}; x_{S_2}, x_{T_2} ignored.
- `Imputation`: full x_S; missing x_{T_2} imputed.
- Two divergences for aligning distributions:
  - $\mathcal{H}$-divergence
  - Wasserstein distance

Datasets and Metrics

- digits (missing half pixels): accuracy

Experimental setting

Baselines

- `Full`: full x_S and x_T.
- `ZeroImputation`: full x_S; missing x_{T_2} set to 0, $x_T = (x_{T_1}, \mathbf{0})$.
- `IgnoreComponent`: only x_{S_1}, x_{T_1}; x_{S_2}, x_{T_2} ignored.
- `Imputation`: full x_S; missing x_{T_2} imputed.
- Two divergences for aligning distributions:
  - $\mathcal{H}$-divergence
  - Wasserstein distance

Datasets and Metrics

- digits (missing half pixels): accuracy
- Amazon product reviews (missing half embeddings): accuracy

Experimental setting

Baselines

- `Full`: full x_S and x_T.
- `ZeroImputation`: full x_S; missing x_{T_2} set to 0, $x_T = (x_{T_1}, \mathbf{0})$.
- `IgnoreComponent`: only x_{S_1}, x_{T_1}; x_{S_2}, x_{T_2} ignored.
- `Imputation`: full x_S; missing x_{T_2} imputed.
- Two divergences for aligning distributions:
  - $\mathcal{H}$-divergence
  - Wasserstein distance

Datasets and Metrics

- digits (missing half pixels): accuracy
- Amazon product reviews (missing half embeddings): accuracy
- challenging real-world advertising datasets[1]: cross-entropy

[1] http://labs.criteo.com/2014/02/kaggle-display-advertising-challenge-dataset/

Results - Target accuracy (↑) and Cross-Entropy (↓)

Dataset	MNIST → USPS		USPS → MNIST		SVHN → MNIST		MNIST → MNIST-M		ads-kaggle	ads-real
Model w/o $\mathscr{R}$	ADV	OT	ADV	OT	ADV	OT	ADV	OT	ADV	ADV
Source-Full	71.5±2.7		74.2±2.7		58.1±1.1		28.3±1.4		NA	
Adaptation-Full	85.8±3.2	92.6±1.7	94.6±2.1	93.9±0.6	78.0±3.4	76.1±1.4	60.8±3.8	46.9±3.9	NA	
Source-ZeroImputation	25.7±3.7		39.2±2.6		31.5±2.		14.4±1.1		0.545±0.019	0.663±0.011
Adaptation-ZeroImputation	48.4±4.8	60.9±6.3	67.5±2.2	65.3±5.2	47.1±5.7	37.5±6.2	34.7±2.5	20.2±2.5	0.397±0.0057	0.660±0.025
Source-IgnoreComponent	52.9±9.7		54.3±1.6		44.6±1.9		19.1±2.6		0.406±0.00046	0.622±0.0048
Adaptation-IgnoreComponent	71.5±3.2	64.0±5.0	80.0±1.4	72.0±1.8	45.5±1.9	47.9±1.8	29.4±1.6	26.8±4.4	0.403±0.0030	0.634±0.0082
Adaptation-Imputation	**74.2±2.3**	**66.8±1.3**	**81.4±0.8**	**72.5±2.7**	**53.8±1.4**	**49.2±1.5**	**57.9±2.3**	**29.2±1.4**	**0.389±0.014**	**0.583±0.013**

Dataset	DVD → Electronics	Books → Kitchen	Kitchen → Electronics	DVD → Books
Source-Full	69.57	73.04	77.88	71.95
Adaptation-Full	73.62	74.09	79.63	72.65
Source-ZeroImputation	58.51	60.52	66.27	61.15
Adaptation-ZeroImputation	64.51	61.08	68.02	62.80
Source-IgnoreComponent	60.21	62.03	67.62	64.35
Adaptation-IgnoreComponent	61.02	64.08	68.47	66.00
Adaptation-Imputation	**72.57**	**72.69**	**78.18**	**72.61**

Conclusion

Our model improves representative baselines:

- on all our datasets
- for two alignment divergences

Ablation studies - Model modules

Ablation study	ADV Model	MNIST → USPS	USPS → MNIST	SVHN → MNIST	MNIST → MNIST-M	ads-kaggle
L_2+L_3 vs. $L_1+L_2+L_3$	$L=\lambda_2 L_2+\lambda_3 L_3$	64.2±1.8 (-13%)	51.3±2.5 (-37%)	44.5±1.4 (-17%)	24.1±2.6 (-58%)	0.410±0.0020 (-5.4%)
ADV-MSE weighting in L_2	$L_2=L_{MSE}$	71.9±3.7 (-3.1%)	**81.4±1.2 (0%)**	52.5±3.7 (-2.4%)	56.5±2.8 (-2.4%)	0.400±0.0014 (-2.8%)
	$L_2=L_{ADV}$	28.6±3.2 (-61%)	39.4±5.2 (-52%)	28.8±3.8 (-46%)	30.0±3.7 (-48%)	0.469±0.13 (-21%)
	$L_2=L_{ADV}+0.005\times L_{MSE}$	47.8±3.7 (-36%)	49.6±5.8 (-39%)	46.0±2.6 (-15%)	50.6±2.2 (-13%)	**0.389±0.014 (0%)**
	$L_2=L_{ADV}+L_{MSE}$	**74.2±2.3 (0%)**	**81.4±0.8 (0%)**	**53.8±1.4 (0%)**	**57.9±2.3 (0%)**	0.401±0.0014 (-3.1%)

Ablation study	ADV Model	DVD → Electronics	Books → Kitchen	Kitchen → Electronics	DVD → Books
ADV-MSE weighting in L_2	$L_2=L_{MSE}$	71.47 (-1.5%)	71.39 (-1.8%)	77.58 (-0.77%)	72.02 (-0.81%)
	$L_2=L_{ADV}+L_{MSE}$	**72.57 (0%)**	**72.69 (0%)**	**78.18 (0%)**	**72.61 (0%)**

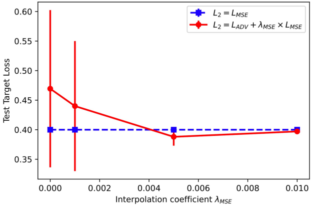


Figure 1: Adaptation-Imputation T CE (↓) on ads-kaggle wrt λ_{MSE}

Conclusion

- L_1 is useful.
- L_{ADV} in L_2 is useful.

Conclusion

Conclusion

Problem

New end-to-end approach for non-stochastic missing data based on an adaptation-imputation problem.

Conclusion

SCIENCES SORBONNE UNIVERSITÉ

Problem

New end-to-end approach for non-stochastic missing data based on an adaptation-imputation problem.

Theory

Clear assumptions and upper-bounds minimized by our model.

Conclusion

Problem

New end-to-end approach for non-stochastic missing data based on an adaptation-imputation problem.

Theory

Clear assumptions and upper-bounds minimized by our model.

Experiments

Superior performance over representative baselines on real-world datasets with extremely different characteristics.

Conclusion

Thank you for your attention !

Code: `https://github.com/mkirchmeyer/adaptation-imputation`

Contact information:

- Matthieu Kirchmeyer: `matthieu.kirchmeyer@gmail.com`

References

References I

Ben-David, Shai et al. (2010). "A theory of learning from different domains". In: *Machine Learning* 79.1, pp. 151–175.

Rubin, Donald B. (Dec. 1976). "Inference and missing data". In: *Biometrika* 63.3, pp. 581–592.