Domain Generalization via Selective Consistency Regularization for Time Series Classification

Paper ID: 303

Institute for Infocomm Research
I²R

Wenyu Zhang[1], Mohamed Ragab[1,2], Chuan-Sheng Foo[1,2]

[1] Institute for Infocomm Research, A*STAR, Singapore
[2] Centre for Frontier AI Research, A*STAR, Singapore

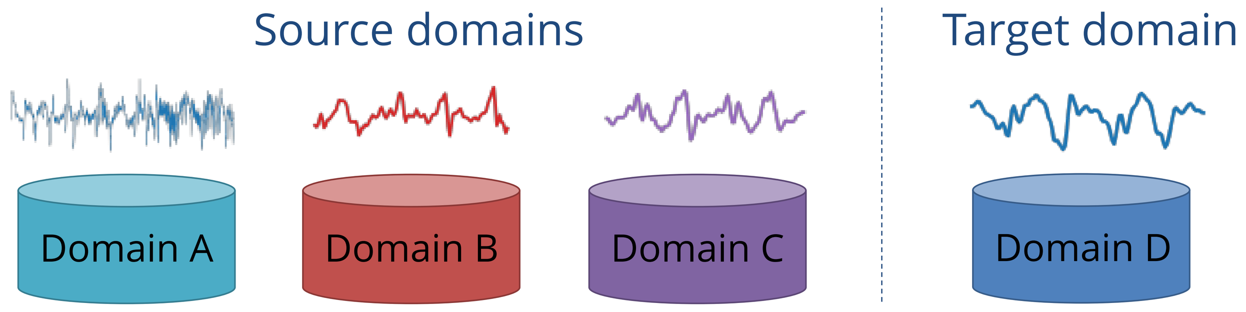
Domain Generalization Problem

Problem of Deep Learning

Deep learning has widely acclaimed performance in various applications. Yet, its success is typically reliant on the assumption that train (source domains) and test (target domain) data are sampled from the same distribution. Many real-world environments are highly dynamic, making test samples to be out of distribution.

Domain Generalization

Domain generalization aims to realize more practical and robust models deployable in the wild. Multiple source domains are leveraged to directly generalize to unseen domains with new data distributions. Domain generalization in time series classification has important applications, but limited existing literature and evaluation.



Flaw of Existing Domain Alignment Approach

- A core strategy is to align all source domains in a common representation space.
- Assumption to learn "domain-invariant" features: For every domain pair $d^{(i)}$ and $d^{(j)}$, for every $x^{(d^{(i)})}$, $\exists\, x^{(d^{(j)})}$ such that $P_i\left(Y\middle|x^{(d^{(i)})}\right) = P\left(Y\middle|f(x^{(d^{(i)})};\Theta)\right) = P\left(Y\middle|f(x^{(d^{(j)})};\Theta)\right) = P_j\left(Y\middle|x^{(d^{(j)})}\right)$ where $f(x;\Theta)$ are features.
- When assumption is invalid, aligning implies that a $x^{(d^{(j)})}$ where $P_i\left(Y\middle|x^{(d^{(i)})}\right) \neq P_j\left(Y\middle|x^{(d^{(j)})}\right)$ will be aligned to $x^{(d^{(i)})}$.

Proposed Method

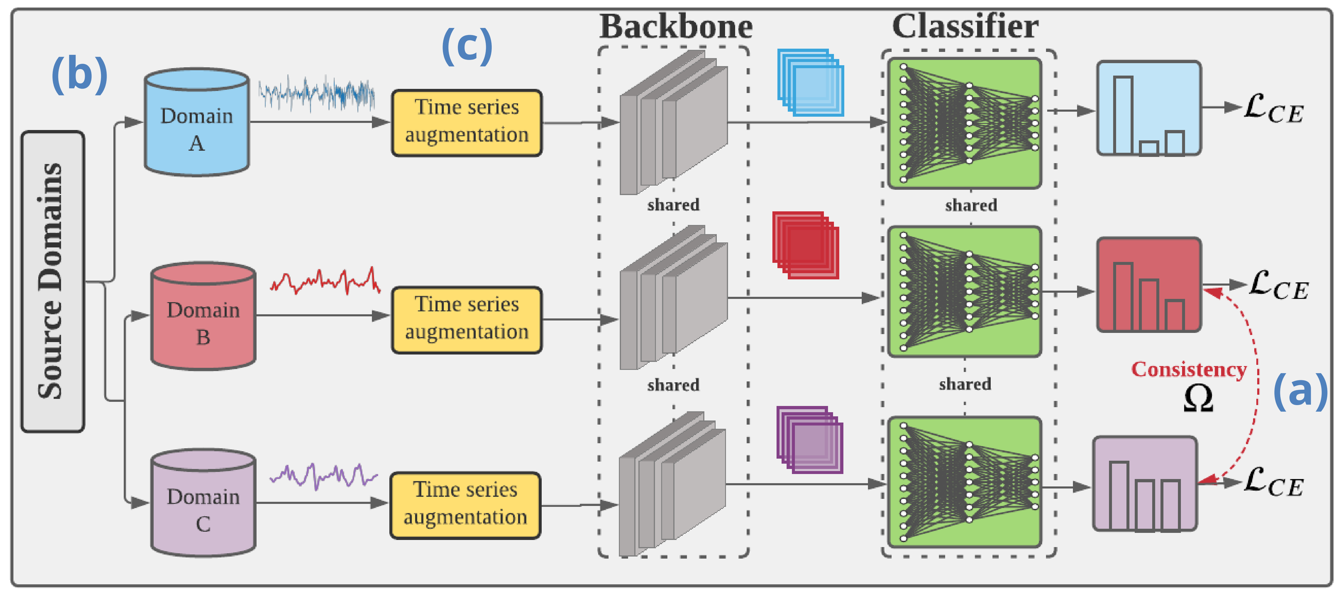
Main Contributions

- Weakens the assumption in existing domain alignment approach by considering inter-domain relationships and selectively enforcing prediction consistency between closely-related source domains.
- Model-agnostic and easy to implement with data augmentation and logit regularization on top of empirical risk minimization.
- Demonstrates better or competitive classification accuracy and calibration compared to existing methods.

Setup

For each source domain d, samples are drawn from a domain-dependent distribution $(x, y) \sim P_d(X, Y)$; x is input series and y is one-hot vector of class label in L classes.

Model consists feature extractor f parameterized by Θ that yields features $z = f(x;\Theta)$, and classifier h parameterized by Ψ that yields logits $g = h(z;\Psi)$. Estimated soft labels are $s = softmax(g)$.



Method

Overall objective: $L(\Theta,\Psi) = L_{CE}(\Theta,\Psi) + \lambda\Omega(\Theta,\Psi)$

Cross-entropy task loss *Selective cross-domain consistency regularization*

(a) Selective cross-domain consistency regularization

- Learn model parameters such that class-conditional predictions $P(G|Y) = P(h(f(X;\Theta);\Psi)|Y)$ is invariant for closely-related domains
- $\Omega(\Theta,\Psi) = \sum_{d^{(i)}} \sum_{d^{(j)}} w\left(d^{(i)}, d^{(j)}\right) \sum_{\ell} \left\| \bar{g}^{(d^{(i)},\ell)} - \bar{g}^{(d^{(j)},\ell)} \right\|_2^2$

Domain similarity *Class-conditional domain logit centroid*

(b) Estimating domain similarity w

- Metadata based similarity: use domain descriptions to group domains
- Learned similarity:

$$w_{learned}\left(d^{(i)}, d^{(j)}\right) = \frac{1}{L}\sum_{\ell} \exp\left(\frac{-\left\|\bar{g}^{(d^{(i)},\ell)} - \bar{g}^{(d^{(j)},\ell)}\right\|_2^2}{2\xi^2}\right)$$

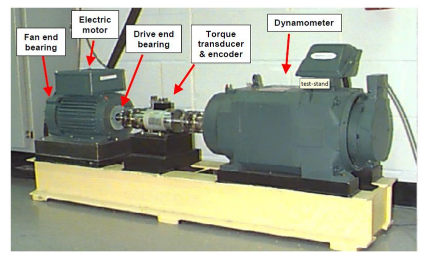
(c) Domain-wise time series augmentation

- Sample augmentation function from pre-built library of augmentations
- Simulates potential test-time domain shifts

Datasets

Bearings Fault Detection (vibration sensor signals)

- 10 classes: 9 faulty classes, 1 healthy class
- 8 domains: 4 loading torques x 2 bearing locations



Human Activity Recognition (device motion sensor signals)

- 6 human activity classes
- 12 domains: 3 users x 4 phone models

MIMIC-III Mortality Prediction (vital signs)

- 2 classes
- 4 domains: 4 age groups

Experimental Results

Generalization Performance:

- Classification performance: accuracy
- Model calibration: expected calibration error (ECE)

$$ECE = \sum_{j=1}^{J} \frac{\|B_j\|}{N} |acc(B_j) - conf(B_j)|$$

for $J = 15$ bins $B_j = \left\{n \middle| n \in \{1,\ldots,N\}, \max(s_n) \in \left(\frac{j-1}{J}, \frac{j}{J}\right]\right\}$ containing indices of samples whose confidence for the predicted class falls in the corresponding interval, with $acc(B_j) = \frac{1}{|B_j|}\sum_{n\in B_j} 1(\hat{y}_n = y_n)$ and $conf(B_j) = \frac{1}{|B_j|}\sum_{n\in B_j} \max(s_n)$.

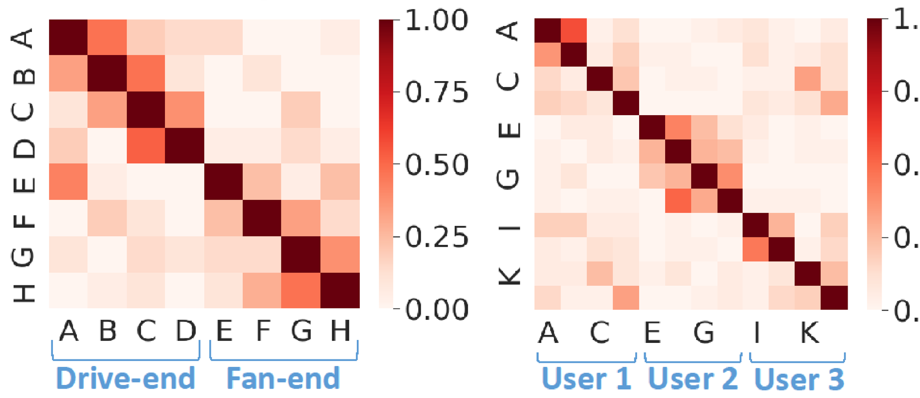
Bearings

Method	Accuracy (%) ↑									ECE (%) ↓
	A	B	C	D	E	F	G	H	Avg	Avg
ERM	65.4	93.8	96.0	71.2	68.4	83.0	94.0	86.2	82.2	14.2
IRM	60.7	87.0	89.2	76.0	62.8	80.9	92.8	88.5	79.7	15.9
GroupDRO	55.7	70.0	77.8	74.8	60.7	59.2	65.6	50.5	64.3	31.9
Interdomain Mixup	62.0	86.5	96.8	76.0	82.0	95.4	97.7	87.2	85.4	13.8
MLDG	62.8	77.6	85.9	72.8	63.3	58.5	60.7	55.0	67.1	30.8
MTL	35.3	64.5	66.6	48.3	47.6	48.2	36.7	44.9	49.0	44.4
Correlation	46.5	79.0	90.0	69.1	71.5	85.5	80.5	83.9	75.7	18.2
VREx	63.8	90.5	97.2	81.6	70.3	83.9	92.0	84.3	83.0	14.3
RSC	62.4	94.4	98.0	86.5	73.4	87.4	97.3	85.7	85.6	12.0
DANN-DG	56.2	84.7	92.2	80.2	70.0	79.1	89.1	90.5	80.3	15.1
CDANN-DG	56.0	80.8	94.8	80.2	70.7	81.4	90.3	84.0	79.8	15.7
CORAL-DG	62.5	77.9	90.0	76.0	63.1	79.2	74.5	83.0	75.8	19.7
MMD-DG	53.9	67.7	84.2	67.6	63.2	74.3	74.7	56.7	67.8	28.7
Ours (Metadata sim.)	86.8	95.3	97.6	79.8	77.4	82.7	93.4	90.8	87.9	9.2
Ours (Learned sim.)	89.1	97.9	97.1	75.8	81.5	85.3	94.4	91.8	**89.1**	**6.2**

Human Activity Recognition

Method	Accuracy (%) ↑													ECE (%) ↓
	A	B	C	D	E	F	G	H	I	J	K	L	Avg	Avg
ERM	86.4	91.6	81.0	91.7	71.3	96.9	96.4	85.9	85.0	88.1	86.6	89.5	87.5	6.0
RSC	87.3	90.5	84.4	92.2	73.9	96.7	96.9	86.2	86.8	87.5	88.5	90.1	88.4	5.3
DANN-DG	84.7	89.5	72.4	92.8	71.2	95.1	94.8	84.2	81.6	84.3	84.9	86.7	85.2	7.5
CDANN-DG	85.6	86.0	79.8	89.6	72.4	93.6	95.6	83.0	81.3	87.0	83.4	85.8	85.2	8.2
CORAL-DG	80.5	76.8	58.4	74.3	62.5	77.5	85.8	74.2	86.8	79.7	86.2	69.1	76.0	15.4
MMD-DG	81.9	74.0	52.9	75.4	60.3	76.8	79.0	74.6	86.9	85.6	83.7	68.9	75.0	17.2
Ours (Metadata sim.)	87.4	91.0	80.7	94.6	75.7	96.5	97.1	86.2	85.0	89.2	89.1	89.8	88.5	**4.1**
Ours (Learned sim.)	87.3	90.6	85.3	93.5	76.0	96.1	96.7	86.0	85.1	88.1	88.6	88.9	88.5	4.9
RSC														
+ Ours (Metadata sim.)	87.2	89.9	86.0	93.6	75.6	96.6	96.4	86.1	85.7	88.6	90.4	90.6	**88.9**	4.5
+ Ours (Learned sim.)	86.3	89.0	84.3	93.3	74.4	96.8	96.9	86.8	87.0	86.7	90.4	89.2	88.5	4.8

Visualization of Learned Domain Relationships



Fraction of runs domain j is the nearest neighbor of domain i

Ablation Study:

Strategy		Avg Acc (%)	
Reg	Aug	Bearings	HHAR
✗	✗	82.2	87.5
✗	✓	86.5	88.1
(Metadata sim.)			
✓	✗	87.1	88.5
✓	✓	87.9	88.5
(Learned sim.)			
✓	✗	86.8	88.3
✓	✓	89.1	88.5

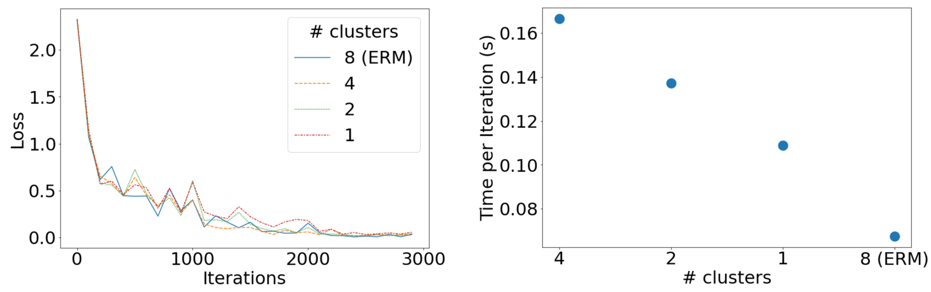
Further Analysis with Different Cluster Assignments:

Classification Accuracy

Cluster assignment	# clusters	Avg Acc (%)
{A},{B},{C},{D},{E},{F},{G},{H}	8 (ERM)	82.2
{A,B},{C,D},{E,F},{G,H}	4	84.4
{A,B,C,D},{E,F,G,H}	2	**87.1**
{A,B,C,D,E,F,G,H}	1	85.6

Good cluster assignments are critical for good performance.

Loss and Average Computation Time



Training loss converges and computation time per iteration is low.

References

[1] https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-universitybearing-data-center-website
[2] http://archive.ics.uci.edu/ml/datasets/heterogeneity+activity+recognition
[3] A. Johnson, T. Pollard, and M. Roger, "MIMIC-III clinical database," PhysioNet, 2016.
[4] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in ICLR, 2021.
[5] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," in JMLR, 2016.
[6] Z. Huang, H. Wang, E.P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in ECCV 2020.