Lifeisgood: Learning Invariant Features via In-Label Swapping for Generalizing Out-of-Distribution in Machine Fault Diagnosis

Zhenling Mo, Zijun Zhang, and Kwok-Leung Tsui

March 14, 2025

To facilitate reading the paper, abbreviations, general rule of superscripts and subscripts, and important symbols (all sorted primarily in alphabet order) are listed as follows:

Nomenclature

Abbreviations

1D-CNN one dimensional convolutional neural network

CaSN causal representation of sufficiency and necessity

CausIRL_MMD causality based invariant representation learning

CondCAD conditional contrastive adversarial domain

CWRU Case Western Reserve University

DG domain generalization

EPdM Ecole Polytechnique de Montr´eal

ERM empirical risk minimization

IB_IRM information bottleneck invariant risk minimization

iDAG invariant directed acyclic graph

IIB invariant information bottleneck

IRM invariant risk minimization

KAIST Korea Advanced Institute of Science and Technology

Lifeisgood learning invariant features via in-label swapping for generalizing out-of-distribution

MDGCML multisource domain-class gradient coordination meta-learning

RDM risk distribution matching

RIDG rational invariance for domain generalization

S0-1 loss swapping 0-1 loss

SAGM sharpness-aware gradient matching

SCE loss swapping cross-entropy loss

SCEDUB swapping cross-entropy difference upper bound

SCIRM sparsity constraint invariant risk Minimization

SelfReg self-supervised contrastive regularization

T-SNE distributed stochastic neighbor embedding

UBFC University of Bourgogne Franche-Comt´e

VNE von Neumann entropy

VREx variance of risk extrapolation

w.r.t with respect to

General Rule of Superscripts and Subscripts (without otherwise specified)

V^2 variable V is squared

V^c variable V is associated with the label c

V^j variable V is indexed by j where $j = 1, 2, 3, ...$

V_i^j variable V is indexed by or associated with both i and j

V^* variable V is the optimal value

V^δ variable V is associated with the swapping set δ

$V^{\mathcal{U}(lb,ub)}$ variable V has a power of $\mathcal{U}$ that is drawn uniformly at random from the range $[lw, up]$ where lb and ub is the lower bound and upper bound respectively

V_i variable V is indexed by i where $i = 1, 2, 3, ...$

V_n variable V is associated with the n-th data instance X_n

$V_n(e_i)$ variable V is associated with the n-th data instance X_n from i-th data domain e_i

$V_{\{0,1\}}$ variable V is associated with the 0-1 loss

$V_{\{i,j\}}$	variable V is associated with a feature pair (Z_i, Z_j)
V_{CE}	variable V is associated with the cross-entropy loss
V_{erm}	variable V is associated with empirical risk minimization
V_{irm}	variable V is associated with invariant risk minimization
$V_{n_i(\delta)}$	variable V is associated with swapping set δ and n_i-th data instance X_{n_i} where $i = 1, 2, 3...$ such that X_{n_1}, X_{n_2}, etc., denote different data instances
V_{te}	variable V is associated with the testing domain distribution
V_{tr}	variable V is associated with the training domain distribution
$V_{y_{n_i}}$	variable V the n_i-th data instance and its label y_{n_i} where $i = 1, 2, 3, ...$

Symbols

$:=$	defined as
$[C]$	label index set
$[E]$	domain index set
$[M]$	data index set
$[N]$	sample index set
&	logic "AND"
$\dot{\mathcal{L}}^{\star\delta}_{\{0,1\}}$	Bayes optimal risks on top of invariant features for the S0-1 loss $\mathcal{L}^{\delta}_{\{0,1\}}$
$\dot{\mathcal{L}}^{\star\delta^\star}_{CE}$	Bayes optimal risks on top of invariant features for the SCE loss $\mathcal{L}^{\delta^\star}_{CE}$
α	descending flag
$\bar{\mathcal{F}}$	a non-empty subset of the model hypothesis class $\mathcal{F}$ such that $\forall \bar{f} \in \bar{\mathcal{F}}$, the featurizer $\bar{h}_\theta$ of $\bar{f}$ generates invariant features
β	Keeping rate
δ	instance of Δ
$\Delta := \mathcal{P}([U])$	meta-swapping set, defined as the power set of $[U]$
$\delta^\star$	optimal swapping set
ϵ	acceptable margin
$\hat{\mathcal{L}}^{\delta^\star}_{CE}$	practical SCE loss
$\hat{P}_{te}$	empirical meta-testing distribution

$\hat{P}_{tr}$ empirical meta-training distribution

κ weight decay

λ balancing weight

λ_{IRM} balancing weight of IRM

$\mathbb{I}$ indicator function

$\mathbb{R}^M$ real value set of dimension M, the data space

$\mathbb{R}^U$ real value set of dimension U, the feature space

$\mathbb{R}_+$ positive real values

$\mathbb{V}$ variance

$\mathbf{X}$ collection of data

$\mathcal{A}(r)$ variable only dependent on the Hellinger distance constraint

$\mathcal{B}(r)$ distribution set determined by the Hillinger distance constraint r

$\mathcal{C}(r)$ variable determined by the Hillinger distance constraint r

$\mathcal{D}(r, N, \sigma, \rho)$ variable only depending on the Hillinger distance constraint r, the sample size N, the confidence parameter σ, and the maximum loss difference between features with the different labels ρ

$\mathcal{F}$ model hypothesis class

$\mathcal{L}_{\{0,1\}}$ 0-1 loss

$\mathcal{L}^{\delta}_{\{0,1\}}$ S0-1 loss with the swapping set δ

$\mathcal{L}_{CE}$ cross-entropy loss

$\mathcal{P}_+$ pair sets of same labels

$\mathcal{P}_-$ pair sets of different labels

$\mathcal{Z}$ feature space for theoretical proofs

μ reference probability measure

$\nabla_{\{\omega,\theta\}}\hat{\mathcal{L}}^{\delta^\star}_{CE}$ gradients of the featurizer and classifier

$\odot$ point-wise product

Ω parameter set of the classifier

ω parameter instance of the classifier

ϕ function that is, non-negative, convex, differentiable at 0, and $\phi'(0) \leq 0$

$\psi(\epsilon)$ variable only determined by the acceptable margin ϵ

ρ maximum loss difference between the features with the different labels

σ confidence parameter

Θ parameter set of the featurizer

θ parameter instance of the featurizer

$\tilde{\mathcal{L}}^{\delta}_{\{0,1\}}$ Bayes optimal risks based on all features of training domains for the S0-1 loss $\mathcal{L}^{\delta}_{\{0,1\}}$

$\tilde{\mathcal{L}}^{\delta^\star}_{CE}$ Bayes optimal risks based on all features of training domains for the S0-1 loss $\mathcal{L}^{\delta^\star}_{CE}$

$\tilde{G}^{c}_{\{i,\tilde{j}\}}$ same labeled features with a randomized index order from $G^{c}_{\{i,j\}}$

$\tilde{Z}_{n_1(\delta)}$ swapped feature of the feature Z_{n_1} based on the swapping set δ

Υ union of index sets where the entries of the features are not identical

υ element of the union of index sets Υ

$\{Z_j, y_j\}_{[J(c)]}$ the group of features and labels where the labels are all c. $[J(c)]$ is the index set of these features / labels, i.e., $j \in [J(c)]$

a_i classification accuracy of a trial

A_t average accuracy at task level

$A_{(t,alg)}$ average accuracy of a method alg at task level

A_{alg} average accuracy of a method alg at dataset level

Avg average accuracy at dataset level

B a batch for computing a loss

C^2_N number of combining 2 elements from a set of N elements

$Count$ count of surpassing ERM

$Count_{all}$ total count of surpassing ERM

d_t standard deviation at task level

$d_{(t,alg)}$ standard deviation of a method alg at task level

e_i the i-th domain instance

f learning model

$Freq_{erm}$ frequency of surpassing ERM

g featurizer

$G_i^c = \left(G_{\{i,j\}}^c, \tilde{G}_{\{i,\tilde{j}\}}^c\right)$ pair-wise cosine similarity between two collections of same labled features corresonding to the label c

g_ω linear classifier with parameter ω

$G_{\{i,j\}}^c = \{Z_j, y_j\}_{[J(c)]}$ grouping features by labels where $[J(c)]$ is the index set of the same labeled features with the label as c

H Hellinger distance

h featurizer

h_θ featurizer with parameter θ

l_n instance-wise loss corresponding to data instance X_n and label instance y_n

M number M

N number N

n_i element of the sample index set $[N]$

N_t number of trials on the DG task

N_{all} total number of comparisons on all datasets

$P_1, P_2 \in \mathcal{B}(r)$ P_1, P_2 are the elements of a set of distributions constraint by the Hellinger distance constraint r

p_n^c the probability of predicting the feature Z_n or the data instance X_n to the label $c \in [C]$, which is also c-th entry of p_n

P_{te} meta-testing distribution

P_{tr} meta-training distribution

Q total number of training step and q is the q-th step

R risk function

$S_N^2(\delta)_{\{0,1\}}$ S0-1 difference between a pair of the swapped feature and the original feature, where N, 2, δ, $\{0,1\}$ indicate sample size, square operation, swapping set, 0-1 loss, respectively

w_υ υ-th entry of the linear classifier weight $W_{y_{n_1}}$

$W_{y_{n_1}}$ class-wise weight vector of the linear classifier g_ω corresponding to the label y_{n_1}

X_n the n-th data instance of $\mathbf{X}$

x_n^M	the M-th entry of X_n
Y	collection of labels
Z_n	the feature with the input as X_n
$Z_n(e_i)$	the n-th feature given by the n-th data instance of the i-th domain
$z_{n_1}^u(e_1)$	the u-th entry of a feature Z with the data instance idnex n_1 and the domain index e_1
$l_{\{0,1\}_n}$	instance-wise 0-1 loss
l_{CEn}	instance-wise cross-entropy loss
$n_1(\delta)$	instance index n_1 that is associate with the swapping set δ
$\mathcal{L}_{CE}^{\delta^\star}$	SCE loss with the optimal swapping set $\delta^\star$
B_0	a batch at training