NEURAL INFORMATION PROCESSING SYSTEMS

LAMDA
Learning And Mining from DatA
www.lamda.nju.edu.cn

LOG: Active Model Adaptation for Label-Efficient OOD Generalization

Jie-Jing Shao, Lan-Zhe Guo, Xiao-Wen Yang and Yu-Feng Li

LAMDA Group

Nanjing University

NeurIPS | 2022

Thirty-sixth Conference on Neural Information Processing Systems

General Idea

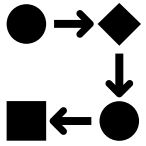
- Recently, Causal Invariant Learning [Arjovsky, 2019] has shown theoretical and empirical effectiveness to deal with Out-Of-Distribution Generalization.
- It still suffer one serious issue:
  - ➢These methods need sufficient multi-source labeled data to remove the source-specific spurious correlation and obtain the generalizable invariant relationship.
- In this work
  - We take a benefit from active learning to address label-efficient generalization.
  - We present an interactive framework LOG, composed of active querying and invariant learning.
  - Theoretical and empirical analysis show quite promising results.

Outline

- Background of OOD Generalization Problem
- Causal Invariant Learning and Its Limitation
- LOG: Label-Efficient OOD Generalization

OOD Generalization

- Machine Learning
  - $P_{tr}(x, y) = P_{te}(x, y)$
- OOD Generalization:
  - $P_{tr}(x, y) \neq P_{te}(x, y)$



College

Factory

Street

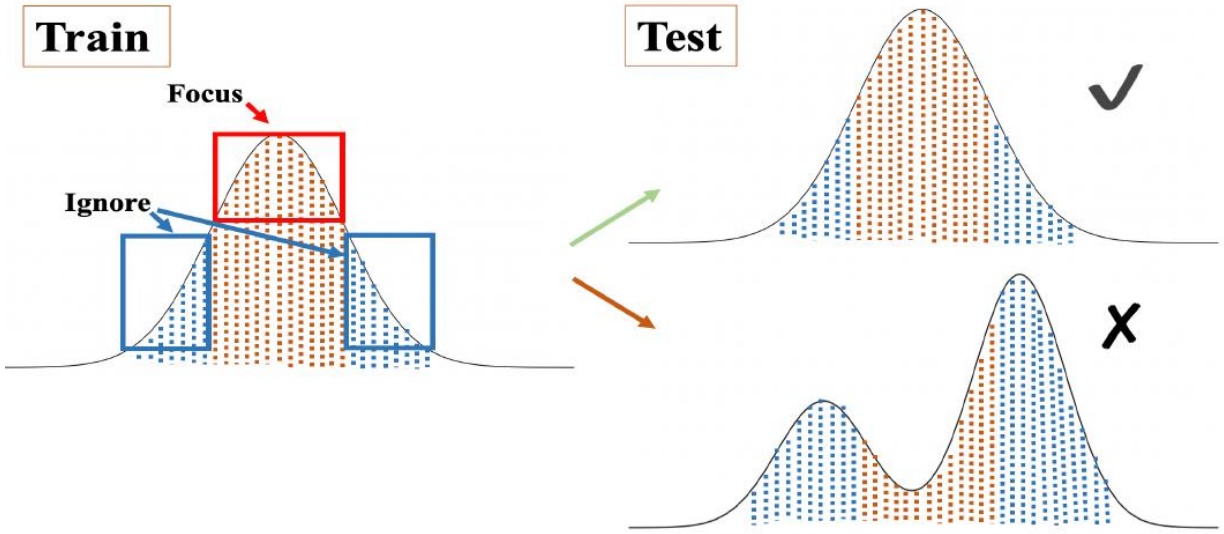
Training data (Campus) -> model Testing data

- Covariate shift
  - $P_{tr}(y|x) = P_{te}(y|x), P_{tr}(x) \neq P_{te}(x)$

Challenge in OOD Generalization

$$\theta_{ERM} = \arg\min_{\theta} \frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} \ell(\theta; X_i; Y_i)$$

- Optimize the average error of data points
- Focus on the major group of data
- Ignore the minor group of data -> Break down under distribution shift



Covariate shift
$P_{tr}(x) \neq P_{te}(x)$

OOD Generalization: A Case

www.lamda.nju.edu.cn

	Class 0 (Car)			Class 1 (Boat)		
Training	X_1	X_2	X_3	X_1	X_2	X_3
45%	Car	Street	Sunny	Boat	Lake	Cloudy
45%	Car	Street	Cloudy	Boat	Lake	Sunny
10%	Car	Lake	Cloudy	Boat	Street	Sunny
Testing	X_1	X_2	X_3	X_1	X_2	X_3
	Car	Lake	Sunny	Boat	Street	Cloudy

ERM will overfit on the task-agnostic training environment.

OOD Generalization: A Case

LAMDA
Learning And Mining from DatA
www.lamda.nju.edu.cn

	Class 0 (Car)			Class 1 (Boat)		
Training	X_1	X_2	X_3	X_1	X_2	X_3
45%	Car	Street	Sunny	Boat	Lake	Cloudy
45%	Car	Street	Cloudy	Boat	Lake	Sunny
10%	Car	Lake	Cloudy	Boat	Street	Sunny
Testing	X_1	X_2	X_3	X_1	X_2	X_3
	Car	Lake	Sunny	Boat	Street	Cloudy

To maintain the performance on unseen testing distribution, we should leverage the latent heterogeneity in data and develop more rational risk minimization approach to achieve Majority Good and Minority Good, resulting in **robustness in varying environments**.

Invariant Learning

- **Assumption (Invariance Assumption)[Based on covariate shift]**

  There exists random variable $\Phi^*(X)$ such that the following properties hold:

  – Invariance property: for all $e_1, e_2 \in \text{supp}(\mathcal{E})$, we have

  $$P^{e_1}(Y|\Phi^*(X)) = P^{e_2}(Y|\Phi^*(X))$$

  – Sufficiency property: $Y = f(\Phi^*) + \epsilon, \epsilon \perp X$

Class 0 (Car)		
X_1	X_2	X_3
Car	Street	Sunny
Car	Street	Cloudy
Car	Water	Cloudy

$$\Phi^*(X) = \mathrm{X}_1$$

$$\mathrm{P}^{e_1}(Y|X_1 = Car) = \mathrm{P}^{e_2}(Y|X_1 = Car)$$
$$X_1 = Car \rightarrow Y = Car$$

To obtain the invariant predictor $\Phi^*(X)$, one can seeks for the Maximal Invariant Predictor, which is defined as follows:

Invariant Learning

- **Assumption (Invariance Assumption)**

  There exists random variable $\Phi^*(X)$ such that the following properties hold:

  - Invariance property: for all $e_1, e_2 \in \text{supp}(\mathcal{E})$, we have
    $$P^{e_1}(Y|\Phi^*(X)) = P^{e_2}(Y|\Phi^*(X))$$
  - Sufficiency property: $Y = f(\Phi^*) + \epsilon, \epsilon \perp X$

- **Definition (Invariant Features & Maximal Invariant Predictor)**

  - The set of invariant features $\mathcal{I}$ with respect to $\mathcal{E}$ is defined as:
    $$\mathcal{I}_{\mathcal{E}} = \{\Phi(X): Y \perp \mathcal{E}|\Phi(X)\} = \{\Phi(X): H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}]\}$$
    where $H[]$ is the Shannon entropy of a random variable. The corresponding maximal invariant predictor (MIP) of $\mathcal{I}_{\mathcal{E}}$ is defined as:
    $$\Phi^* = \arg\max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi)$$
  - Where I() measures Shannon mutual information between two random variables

- Generalization: the solution $f^*(X) = P(Y|\Phi^*)$ has $\max_{e \in \mathcal{E}} R(f^*; D^e) \leq \min_f \max_{e \in \mathcal{E}} R(f; D^e)$

Invariant Learning

- Given: $D = \{D^e\}_{e\in\mathcal{E}_{tr}}, D^e = \{(x_i^e, y_i^e)\}$
- Goal: $\arg\min_{f} \max_{e\in\mathcal{E}} L(f|e)$ (worst-case risk)
- Invariant Risk Minimization (Arjovsky & Bottou, 2019)

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w\circ\Phi)$$
$$\text{subject to} \quad w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{\text{tr}}.$$

- IRMv1(Arjovsky & Bottou, 2019)

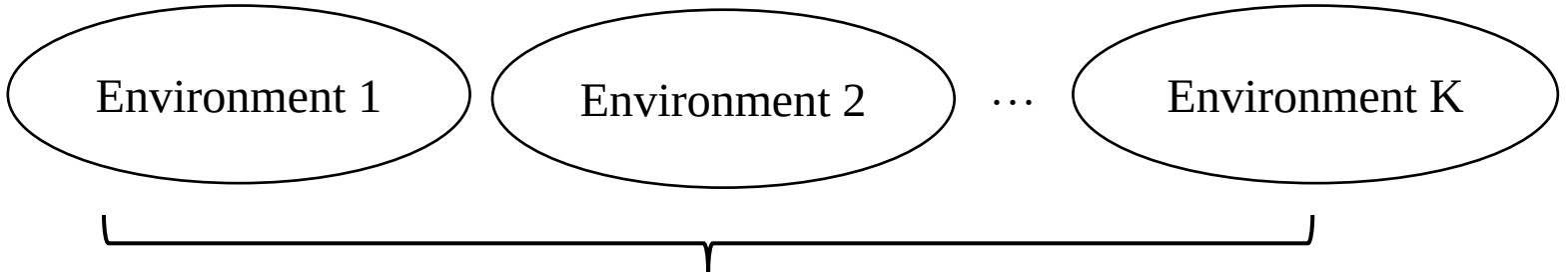
$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda\cdot\|\nabla_{w|w=1.0} R^e(w\cdot\Phi)\|^2,$$

- Variance penalty regularizer (Koyama & Yamaguchi, 2020)

$$\min_{\Phi:\mathcal{X}\to H} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda\cdot\left\|Var_{\mathcal{E}_{tr}}(\nabla_\theta L^e)\circ\Phi\right\|^2$$

Label-sufficient vs. Label-scarce

- Invariant Learning is a typical way to address OOD Generalization, however



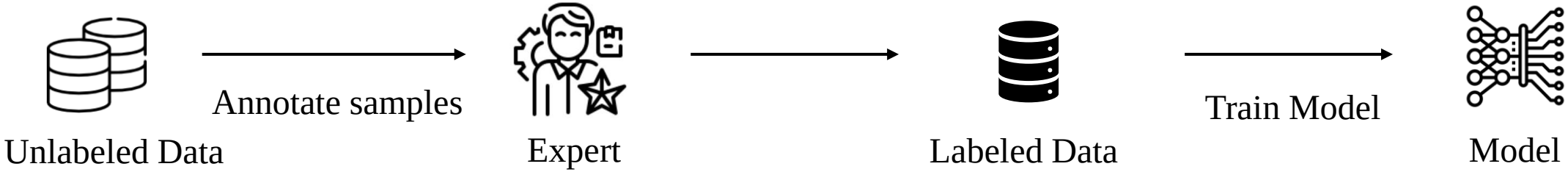
They find the invariant predictor across multiple distributions.

Their robustness rely on the:
multiple training sources and **plenty labeled data**. (label-sufficient scenario)

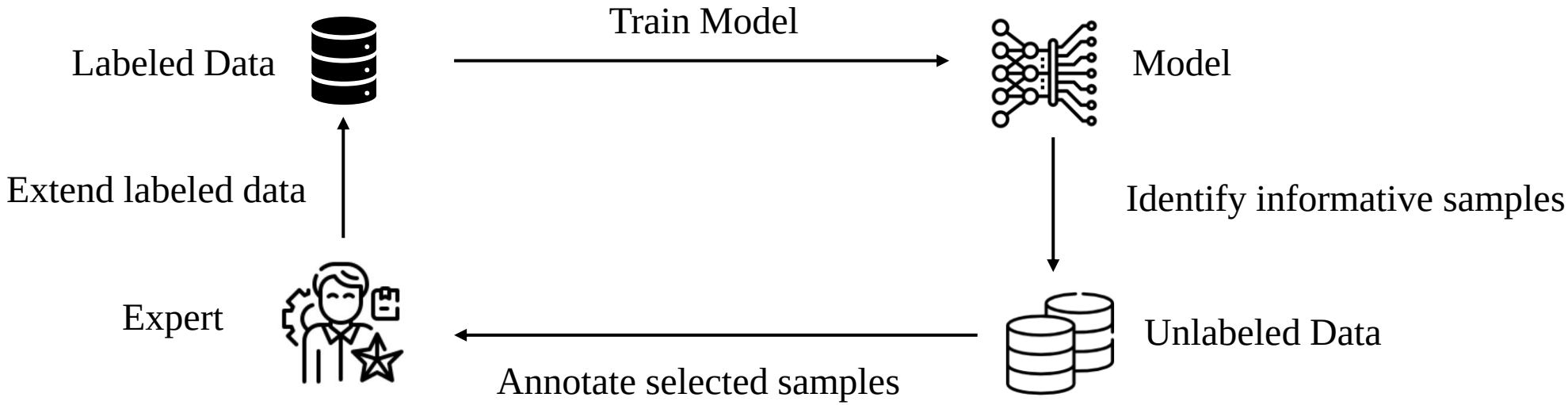
- These strategies may perform poorly in the label-scarce applications.
  - It is desirable to consider label-efficient worst-case robustness.

Passive vs. Active Learning

- Traditional pipeline of Machine Learning

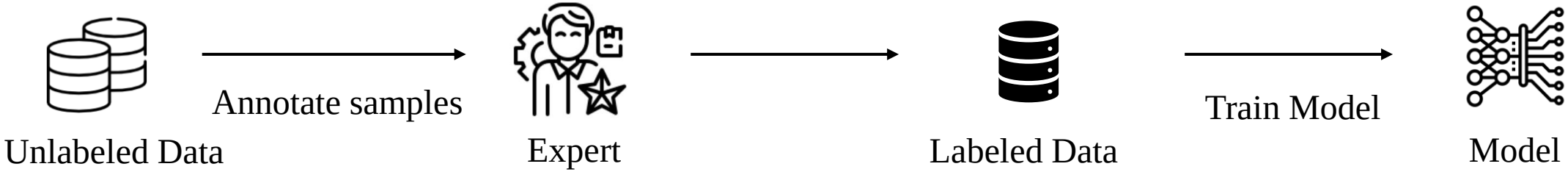


- Active Learning, interactively query the label of samples

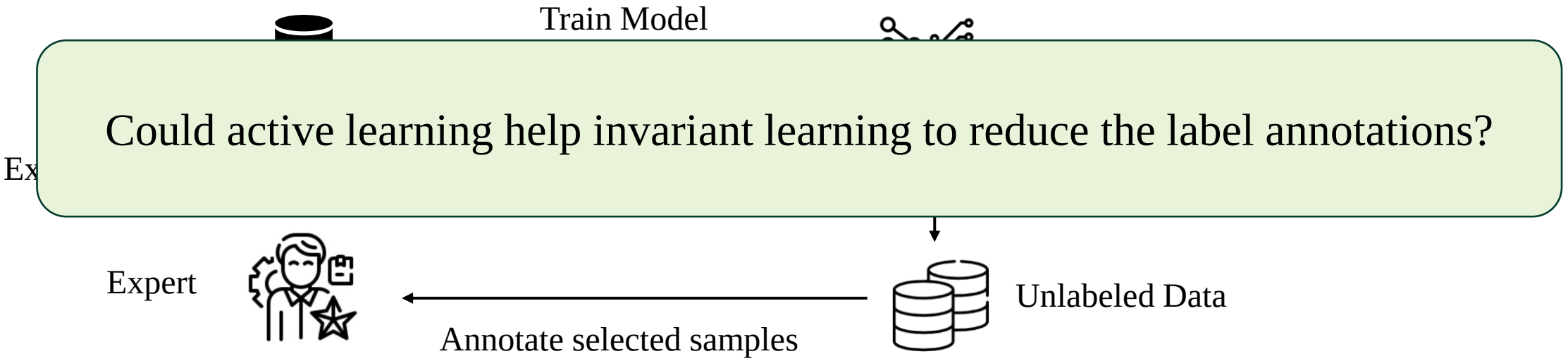


Passive vs. Active Learning

- Traditional pipeline of Machine Learning



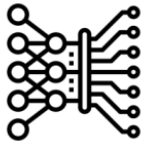
- Active Learning, interactively query the label of samples



Interactive algorithm: LOG

Labeled Data

Improve generalization



Model

Active Annotation

Invariant Learning

Expert

Query [**informative**] samples

Unlabeled Data

Q: Which samples are **informative** for Invariant Learning

Interactive algorithm: LOG

Q: Which samples are **informative** for Invariant Learning

A: The samples which violate the current invariant relationship Φ.

	Class 0 (Car)			Class 1 (Boat)		
Env	X_1	X_2	X_3	X_1	X_2	X_3
1	Car	Street	Sunny	Boat	Lake	Cloudy
2	Car	Street	Cloudy	Boat	Lake	Sunny
3	Car	Lake	Cloudy	Boat	Street	Sunny
4	Car	Street	Sunny	Boat	Lake	Windy

Given D^1, D^2 $\quad \Phi(D^1 \cup D^2) = X_1 \cdot X_2$

Annotate D^3? $\quad \Phi'(D^1 \cup D^2 \cup D^3) = X_1 \quad \surd$

Annotate D^4? $\quad \Phi'(D^1 \cup D^2 \cup D^4) = X_1 \cdot X_2 \quad \times$

Interactive algorithm: LOG

Q: Which samples are **informative** for Invariant Learning

A: The samples which violate the current invariant relationship Φ.

	Class 0 (Car)			Class 1 (Boat)		
Env	X_1	X_2	X_3	X_1	X_2	X_3
1	Car	Street	Sunny	Boat	Lake	Cloudy
2	Car	Street	Cloudy	Boat	Lake	Sunny
3	Car	Lake	Cloudy	Boat	Street	Sunny

How to locate these samples $\Phi(D') \neq \Phi(D_1 \cup D_2)$ from data pool?

Annotate D^3? $\Phi'(D^1 \cup D^2 \cup D^3) = X_1$ $\surd$

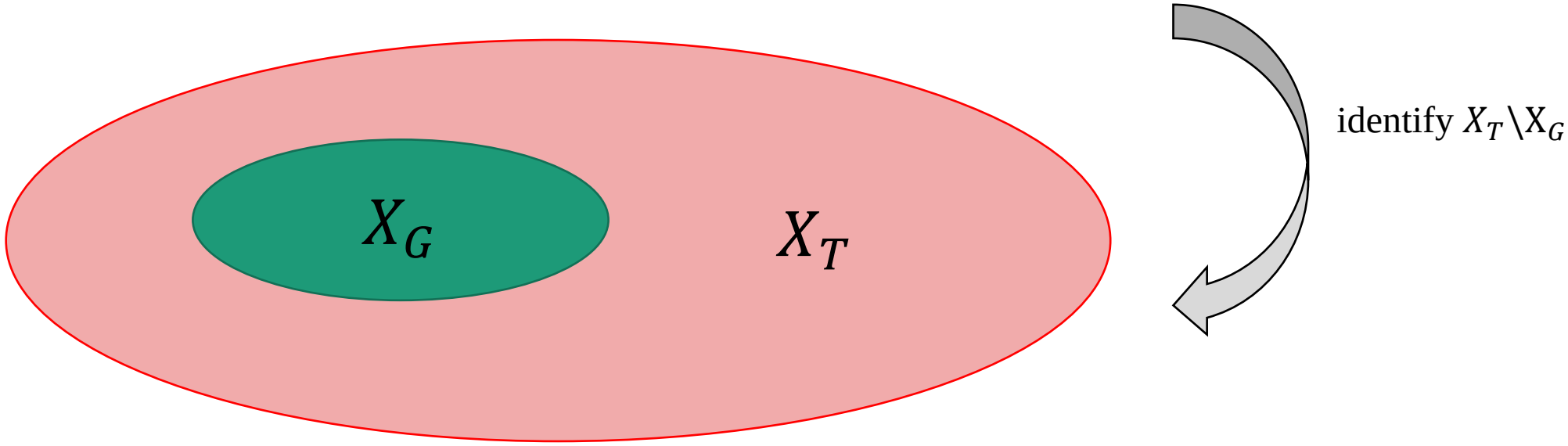
Annotate D^4? $\Phi'(D^1 \cup D^2 \cup D^4) = X_1 \cdot X_2$ $\times$

The Proposed LOG

- Unlabeled data pool could be formulated as:

$$X_T = \theta X_G + (1-\theta) X_U$$

  - where X_G has the same invariance with current model $P\big(\Phi(X_S)\big) = P(\Phi(X_G))$
  - X_U represents the samples which violate the current invariance.
- We build an indicator $g \circ \Phi(x): \mathcal{X} \to \{1, -1\}$: $\mathbb{I}(x \text{ from } X_G)$



The Proposed LOG

- Unlabeled data pool could be formulated as:

$$X_T = \theta X_G + (1-\theta) X_U$$

  - where X_G has the same invariance with current model $P\big(\Phi(X_S)\big) = P(\Phi(X_G))$
  - X_U represents the samples which violate the current invariance.

- We build an indicator $g \circ \Phi(x): \mathcal{X} \to \{1, -1\}$: $\quad \mathbb{I}(x \text{ from } X_G)$
  - With the help of current labeled data, g could be established via risk rewriting.

$$R(g) = \theta E_{x\sim X_G}[\ell(g \circ \Phi(x), 1)] + (1-\theta) E_{x\sim X_U}[\ell(g \circ \Phi(x), -1)]$$
$$= \theta E_{x\sim X_S}[\ell(g \circ \Phi(x), 1) - \ell(g \circ \Phi(x), -1)] + E_{x\sim X_T}[\ell(g \circ \Phi(x), -1)]$$

  - g is a unbiased estimator and converges with a $O(\frac{1}{\sqrt{N_S}} + \frac{1}{\sqrt{N_T}})$ order.

The Proposed LOG

- Interactive framework:
  - **Active Query:**

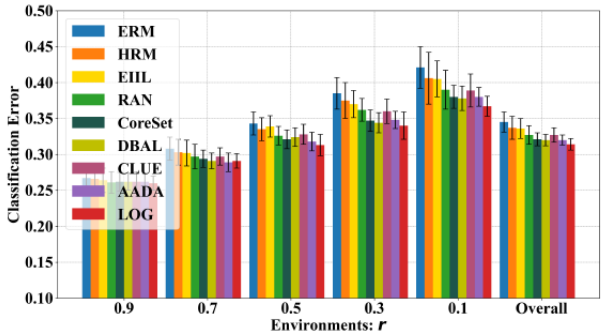
    we actively annotate samples which violate the current invariant relationship Φ.
  - **Invariant Learning**:

    update the Φ on the labeled data $\{D, Q\}$.

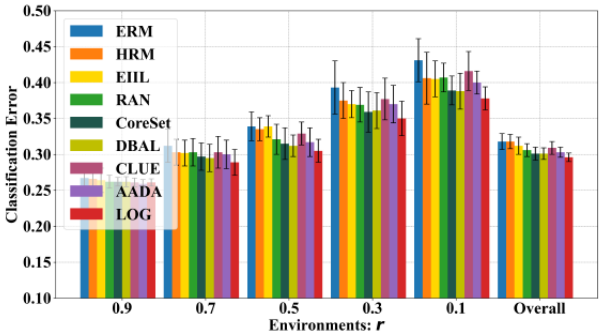
- Efficiency**:** (linear structural causal model case)
  - Each iteration reduces the freedom of Φ.
  - The Φ will converge to the ideal Φ^* at most w steps,

    where w is the freedom of the latent intervention variable.

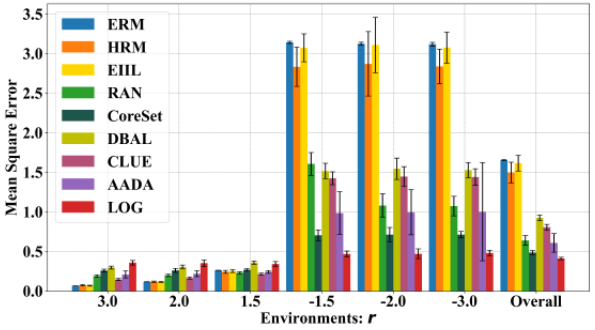
Experimental Results



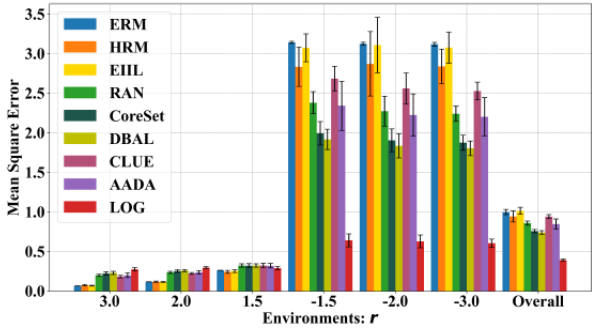
(a) Error under Anti-Causal Effect Shifts



(b) Error under Anti-Causal Effect Shifts (Imbalanced)

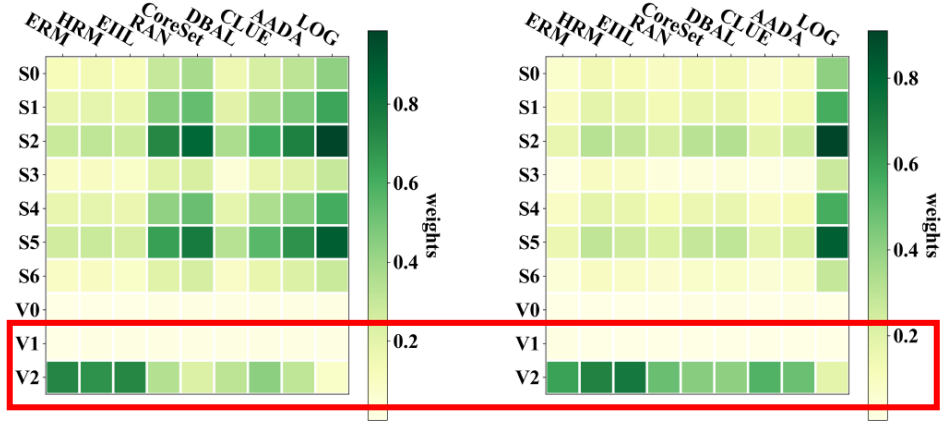


(c) Error under Selection Bias Shifts



(d) Error under Selection Bias Shifts (Imbalanced)

Figure 2: Results on varying base distributions (under 10% labeling budgets).

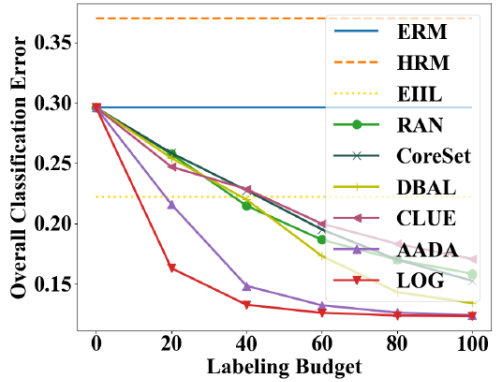


(a) Uniform Case (b) Imbalanced Case

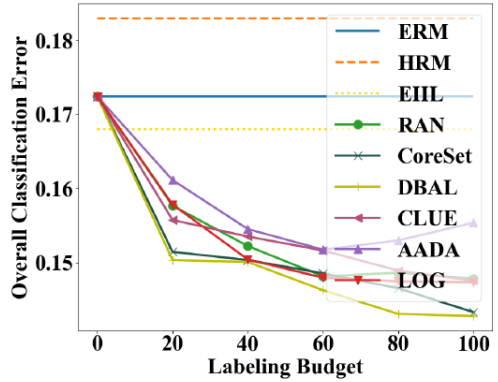
Figure 3: Feature importance for each method.

LOG has shown advantages on both performance and robustness.
LOG clearly reduces overfitting to intervention variables.

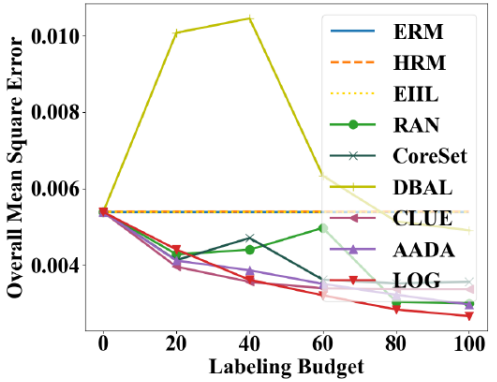
Experimental Results



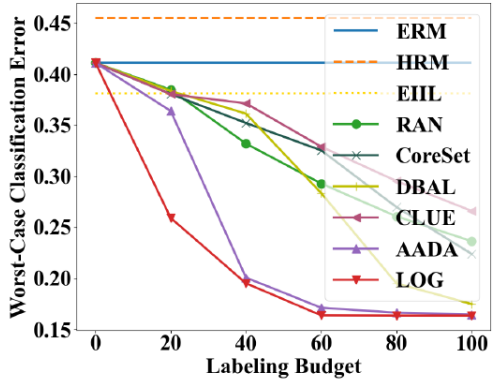
(a) Overall error on Insurance



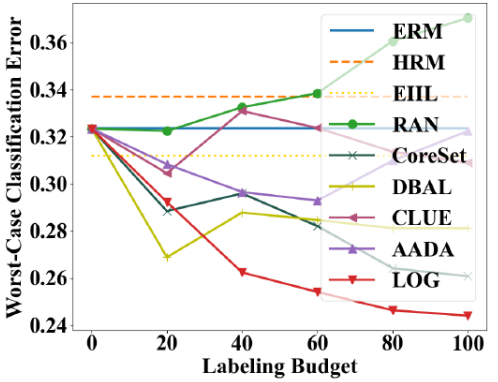
(b) Overall error on Income



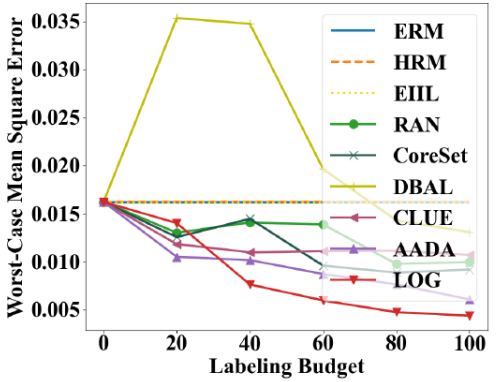
(c) Overall error on House



(d) Robustness error on Insurance



(e) Robustness error on Income



(f) Robustness error on House

The proposed LOG outperforms pervious methods on real-world data with region, person and time shift.

Take-home Message

LAMDA
Learning And Mining from DatA
www.lamda.nju.edu.cn

Active Model Adaptation for Label-Efficient OOD Generalization

- ✓A resource-constrained perspective for OOG generalization.
- ✓An interactive framework LOG with provable convergence.
- ✓Clear effectiveness on a series of datasets.

Thanks for listening!

WeChat

Homepage