Domain Generalization via Multidomain Discriminant Analysis

Shoubo Hu [1] Kun Zhang [2] Zhitang Chen [3] Laiwan Chan [1]

[1] The Chinese University of Hong Kong [2] Carnegie Mellon University [3] Huawei Noah's Ark Lab

Motivation

Background: distribution shift, which is ubiquitous in practice, is the major source of model performance reduction when applied on previously unseen data.

Objective (general)

Incorporate the knowledge from multiple source domains to improve the generalization ability of classifiers on unseen target domains. [1]

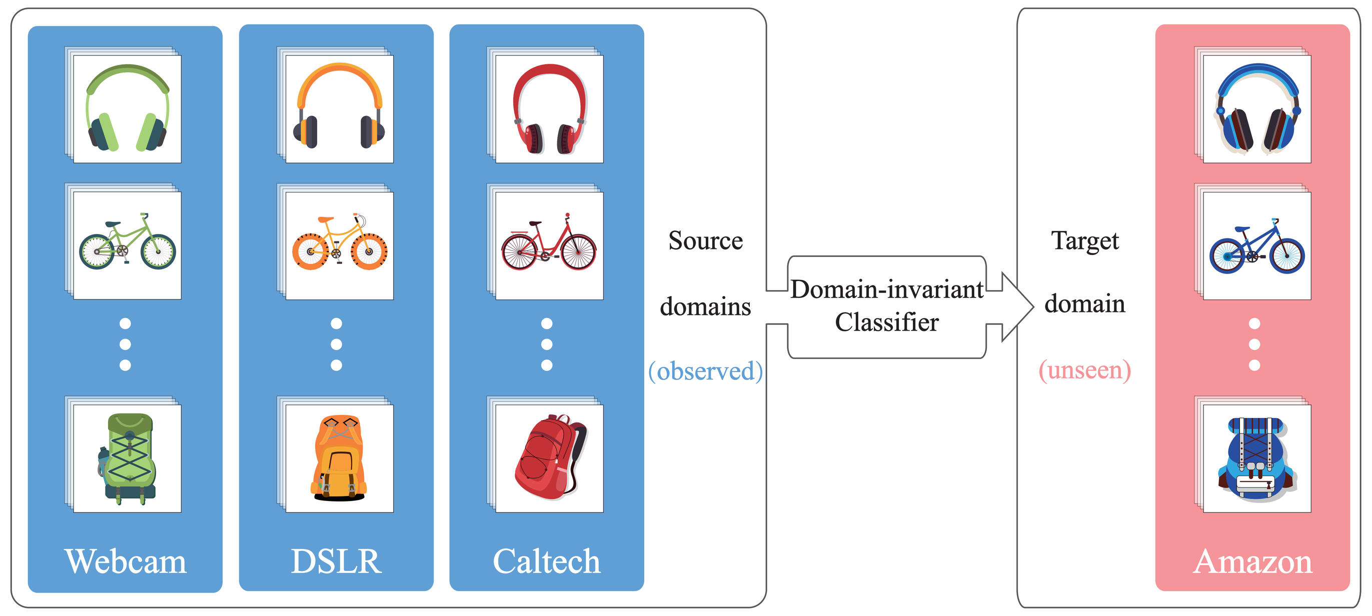

Figure: Illustration of DG on Office+Caltech Dataset. One is given source domains: Webcam, DSLR, Caltech, and aims to train a classifier generalizes well on target domain Amazon, which is unavailable in training.

Problem Setup

Notation	Description	Notation	Description
X, Y	feature/label variable	$\boldsymbol{x}, y$	feature/label instance
m, n	# domains/instances	Z	domain-invariant latent variable
$\mathbb{P}_j^s$	class-conditional distribution	μ_j^s	kernel mean embedding of $\mathbb{P}_j^s$
u_j	mean representation of class j	$\bar{u}$	mean representation of $\mathcal{D}$

Model assumptions

A domain is defined to be a joint distribution $\mathbb{P}(X,Y)$.

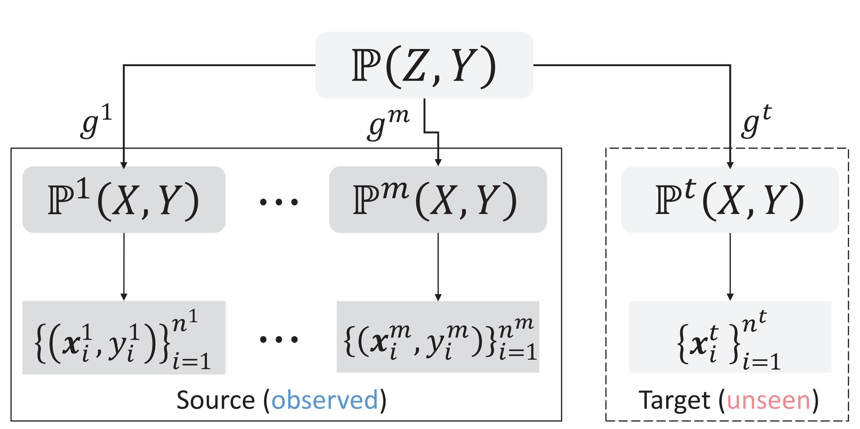

Figure: Domain generalization model assumption. m domains are uniformly sampled from a set of domains and are called the **source domains**. A model trained on m source domains is expected to generalize well on an unseen domain $\mathbb{P}^t(X,Y)$, which is called the **target domain**.

Objective (our method)

We aim to learn a feature transformation, $h(X) : \mathcal{X} \mapsto \mathbb{R}^q$, from the input space to a q-dimensional transformed space $\mathbb{R}^q$ such that

1. source instances of the same class are close in $\mathbb{R}^q$;
2. source instances of different classes are distant in $\mathbb{R}^q$.

Postulate 1. Independence of cause and mechanism [2]

If Y causes X ($Y \to X$), then the marginal distribution of the cause, $\mathbb{P}(Y)$, and the conditional distribution of the effect given the cause, $\mathbb{P}(X|Y)$, are "independent" in the sense that $\mathbb{P}(X|Y)$ contains no information about $\mathbb{P}(Y)$.

According to the postulate above, we factorize the joint distributions in the causal direction

$$\mathbb{P}(X,Y) = \mathbb{P}(Y)\mathbb{P}(X|Y), \tag{1}$$

and manipulate the class-conditional distributions $\mathbb{P}^s(X|Y=j)$ for $s = 1, \dots, m$ and $j = 1, \dots, c$ instead of marginal distributions in most previous works [3].

Regularization Measures

Within-class measures (objective 1)

Average Domain Discrepancy $\Psi^{add} := \frac{1}{c\binom{m}{2}} \sum_{j=1}^{c} \sum_{1 \le s < s' \le m} \|\mu_j^s - \mu_j^{s'}\|_{\mathcal{H}}^2$

Muldidomain within-class scatter $\Psi^{mws} := \frac{1}{n} \sum_{j=1}^{c} \sum_{s=1}^{m} \sum_{i=1}^{n_j^s} \|\phi(\boldsymbol{x}_{i\in j}^s) - u_j\|_{\mathcal{H}}^2$

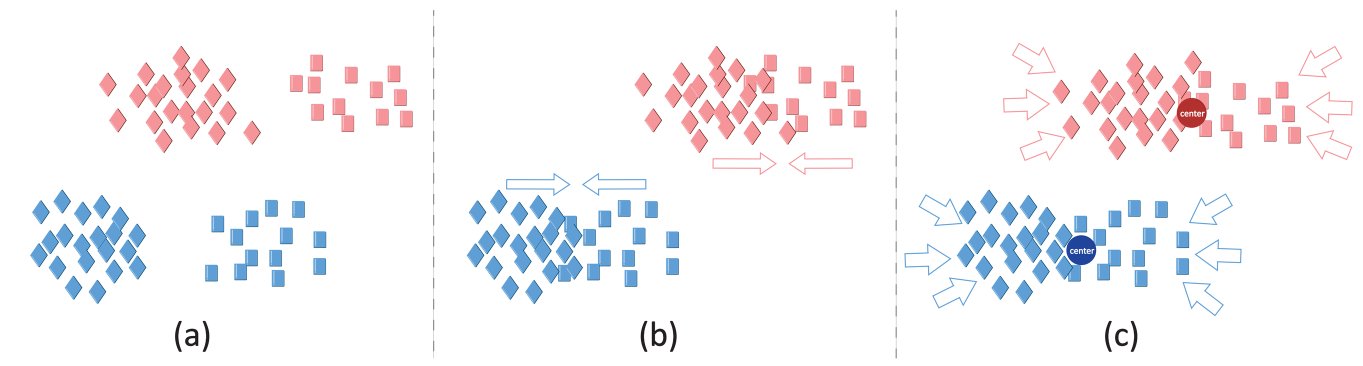

Figure: Illustration of Ψ^{add} and Ψ^{mws} (colors - classes; markers - domains). (a) The distribution of data in the subspace $\mathbb{R}^q$ transformed by some $\mathbf{W}^0$. (b) Minimizing Ψ^{add} makes the means within each class closer. (c) Minimizing Ψ^{mws} makes the distribution of each class more compact towards its mean representation.

Between-class measures (objective 2)

Average class discrepancy $\Psi^{acd} := \frac{1}{\binom{c}{2}} \sum_{1 \le j < j' \le c} \|u_j - u_{j'}\|_{\mathcal{H}}^2$

Multidomain between-class scatter $\Psi^{mbs} := \frac{1}{n} \sum_{j=1}^{c} n_j \|u_j - \bar{u}\|_{\mathcal{H}}^2$

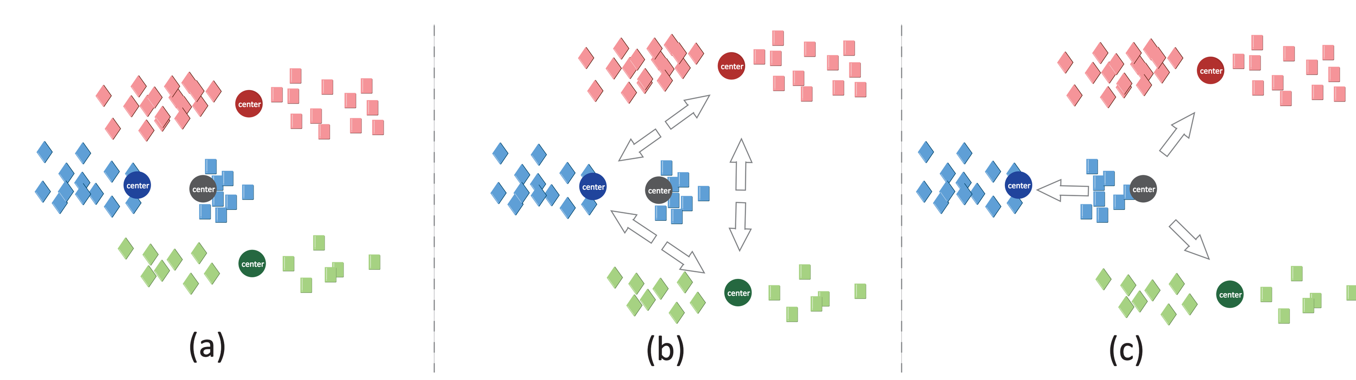

Figure: Illustration of Ψ^{acd} and Ψ^{mbs} (colors - classes; markers - domains). (a) The distribution of data in the subspace $\mathbb{R}^q$ transformed by some $\mathbf{W}^0$. (b) Maximizing Ψ^{acd} makes the distances between each pair of mean representations larger. (c) Maximizing Ψ^{mbs} makes the average distance between the overall mean and the mean representation of different classes larger.

Multidomain Discriminant Analysis

The optimization problem

We unify regularization measures and solve the following optimization problem:

$$\arg\max \frac{\Psi^{acd} + \Psi^{mbs}}{\Psi^{add} + \Psi^{mws}}. \tag{2}$$

We term the proposed method Multidomain Discriminant Analysis (MDA) and summarize the algorithm below

Input: $\mathcal{D} = \{\mathcal{D}^s\}_{s=1}^{m}$ - the set of instances from m domains;
α, β, γ - trade-off parameters;

Feature transformation learning **Output**: Optimal projection $\mathbf{B}_{n\times q}$; corresponding eigenvalues $\mathbf{\Gamma}$.	Target feature transformation **Output**: the transformed target features $\mathbf{X}^t$
• Construct kernel matrix $\mathbf{K}$, whose entry on ith row and i'th column $[\mathbf{K}]_{ii'} = k(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$; • Compute matrices corresponding to regularization measures; • Center the kernel matrix as $\mathbf{K} \leftarrow \mathbf{K} - \mathbf{1}_n\mathbf{K} - \mathbf{K}\mathbf{1}_n + \mathbf{1}_n\mathbf{K}\mathbf{1}_n$, where $\mathbf{1}_n \in \mathbb{R}^{n\times n}$ denotes a matrix with all entries equal to $\frac{1}{n}$; • Solve for the projection $\mathbf{B}$ and corresponding eigenvalues $\mathbf{\Gamma}$, then select q leading components.	• Denote the set of instances from the target domain by $\mathcal{D}^t$, one first constructs the kernel matrix $\mathbf{K}^t$, where $[\mathbf{K}^t]_{i'i} = k(\boldsymbol{x}_{i'}^t, \boldsymbol{x}_i), \forall \boldsymbol{x}_{i'}^t \in \mathcal{D}^t$, $\forall \boldsymbol{x}_i \in \mathcal{D}$; • Center the kernel matrix as $\mathbf{K}^t \leftarrow \mathbf{K}^t - \mathbf{1}_{n^t}\mathbf{K}^t - \mathbf{K}^t\mathbf{1}_n + \mathbf{1}_{n^t}\mathbf{K}^t\mathbf{1}_n$, where n^t is the number of instances in $\mathcal{D}^t$; • Then the transformed features of the target domain are given by $\mathbf{X}^t = \mathbf{K}^t\mathbf{B}\mathbf{\Gamma}^{-\frac{1}{2}}$.

Learning Theory Analysis

Theorem 3. Under assumptions 2 - 4, and assuming that all source sample sets are of the same size, i.e. $n^s = \bar{n}$ for $s = 1, \dots, m$, then with probability at least $1-\delta$ there is

$$\sup_{\|f\|_{\mathcal{H}_{\bar{k}}} \le 1} \left| \frac{1}{m} \sum_{s=1}^{m} \frac{1}{n^s} \sum_{i=1}^{n^s} \ell\left(f(\hat{\tilde{X}}_i^s \mathbf{W}), y_i^s\right) - \mathcal{E}(f,\infty) \right|$$
$$\le U_\ell \left(\sqrt{\frac{\log 2\delta^{-1}}{2m\bar{n}}} + \sqrt{\frac{\log \delta^{-1}}{2m}} \right) + \sqrt{\mathrm{tr}(\mathbf{B}^T\mathbf{K}\mathbf{B})} \left(c_1 \sqrt{\frac{\log 2\delta^{-1} m}{\bar{n}}} + c_2 \left(\sqrt{\frac{1}{m\bar{n}}} + \sqrt{\frac{1}{m}} \right) \right). \tag{3}$$

The first term is of order $O(m^{-1/2})$ and converges to zero as $m \to \infty$. The second term involves the size of the distortion $\mathrm{tr}(\mathbf{B}^T\mathbf{K}\mathbf{B})$ introduced by $\mathbf{B}$. Therefore, a poor choice of $\mathbf{B}$ would loose the guarantee.

Experiments

Synthetic data. Data: two-dimensional Gaussian. Domains: two source domains and one target domain. Classes: three classes in each domain.

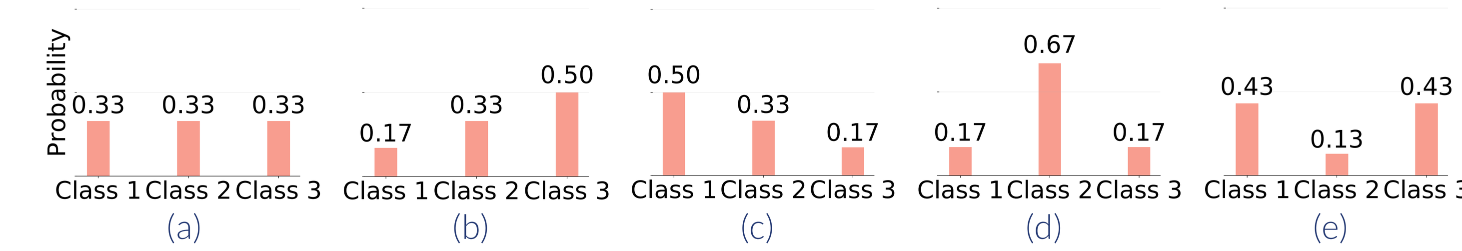

Figure: Class Prior Distributions $\mathbb{P}(Y)$ in Synthetic Experiments.

Table: Accuracy (%) of Synthetic Experiments (**bold red** and **bold** indicate the best and second best).

$\mathbb{P}^1(Y)$ $\mathbb{P}^2(Y)$	(a) (a)	(b) (a)	(c) (a)	(d) (a)	(e) (a)	(a) (b)	(a) (c)	(a) (d)	(a) (e)
SVM	56.00	34.00	33.33	33.33	33.33	33.33	40.00	36.00	60.00
KPCA	66.00	62.00	66.67	33.33	33.33	65.33	36.00	40.00	14.00
KFD	78.67	38.67	46.00	74.67	47.33	49.33	34.00	19.33	76.00
L-SVM	56.00	60.00	64.00	62.00	60.67	64.67	45.33	46.00	59.33
DICA	**93.33**	84.67	76.00	**84.00**	84.67	54.00	**95.33**	71.33	**88.67**
SCA	79.33	72.00	**84.67**	57.33	76.00	59.33	84.67	61.33	81.33
CIDG	90.67	**87.33**	74.67	77.33	**86.67**	**83.33**	**92.00**	**82.00**	86.00
MDA	**96.67**	**96.00**	**97.33**	**94.00**	**94.00**	**91.33**	**95.33**	**94.00**	**94.00**

VLCS Dateset. Data: DeCAF$_6$ features of 4096 dimensions. Domains: V(VOC2007), L(LabelMe), C(Caltech), and S(SUN09). Classes: five classes (bird, car, chair, dog, and person).

Table: Accuracy (%) of VLCS Dataset

Target	V	L	C	S	V, L	V, C	V, S	L, C	L, S	C, S
1NN	60.19	53.57	89.94	55.74	57.26	58.54	50.59	66.06	58.13	66.25
SVM	**68.57**	59.26	**93.99**	**65.27**	**61.80**	**64.39**	**55.89**	70.08	**64.10**	**71.09**
KPCA	60.69	54.86	83.89	55.61	57.54	57.50	49.46	67.48	56.05	66.15
KFD	61.64	**60.54**	86.78	58.75	57.33	46.84	53.20	70.03	61.64	67.87
L-SVM	58.14	39.87	75.56	52.92	52.25	56.64	48.27	61.24	56.65	66.27
CCSA	60.39	58.80	86.88	59.87	59.27	55.02	51.56	69.94	61.41	68.49
DICA	62.71	59.38	86.15	57.28	58.11	55.08	55.17	70.01	61.44	70.30
SCA	62.13	58.24	88.48	**60.66**	**60.66**	57.59	54.66	**71.90**	61.57	70.71
CIDG	64.16	57.91	90.11	59.48	60.54	54.56	55.77	70.74	62.48	69.83
MDA	**66.86**	**61.78**	**92.64**	59.58	59.60	**63.72**	**55.98**	**72.88**	**62.83**	**72.00**

References

[1] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178--2186, 2011.

[2] Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168--5194, 2010.

[3] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 10--18, 2013.