On f-Divergence Principled Domain Adaptation: An Improved Framework

Ziqiao Wang
School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa, Canada
zwang286@uottawa.ca

Yongyi Mao
School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa, Canada
ymao@uottawa.ca

I. Introduction and Background

Unsupervised domain adaptation (UDA) plays a crucial role in addressing distribution shifts in machine learning. Recently, [1] proposed an f-divergence-based domain learning framework. However, their f-divergence-based discrepancy has an unnecessary absolute value function, thus leading to an overestimation of the domain discrepancy. In this work, we introduce a new measure, f-domain discrepancy (f-DD), and give a novel target error bound for UDA.

a) UDA Setup: Let $\mathcal{X}$ and $\mathcal{Y}$ be the input space and the label space. Let $\mathcal{H}=\{h:\mathcal{X}\to\mathcal{Y}\}$ be the hypothesis space. Consider a single-source UDA setting, where μ and ν are two unknown distributions on $\mathcal{X}\times\mathcal{Y}$, characterizing respectively the source and the target domain. Let $\mathcal{S}=\{(X_i,Y_i)\}_{i=1}^n\sim\mu^{\otimes n}$ be a labeled source-domain sample and $\mathcal{T}=\{X_j\}_{j=1}^m\sim\nu^{\otimes m}$ be an unlabelled target-domain sample. We use $\hat{\mu}$ and $\hat{\nu}$ to denote the empirical distributions on $\mathcal{X}$ corresponding to $\mathcal{S}$ and $\mathcal{T}$, respectively. The objective of UDA is to find a hypothesis $h\in\mathcal{H}$ based on $\mathcal{S}$ and $\mathcal{T}$ that "works well" on the target domain. Let $\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}_0^+$ be a symmetric loss. The target error for each $h\in\mathcal{H}$ is defined as $R_\nu(h)\triangleq\mathbb{E}_{(X,Y)\sim\nu}\left[\ell(h(X),Y)\right]$, and the error in the source domain, $R_\mu(h)$, is defined in the same way. Since μ and ν are unknown to the learner, one often uses recourse to the empirical risk in the source domain, which, for a given $\mathcal{S}$, is defined as $R_{\hat{\mu}}(h)\triangleq\frac{1}{n}\sum_{i=1}^n\ell(h(X_i),Y_i)$. We will simply use $\ell(h,h')$ to represent $\ell(h(x),h'(x))$.

b) Background on f-divergence: The family of f-divergence is defined as follows.

Definition I.1. Let P and Q be two distributions on Θ. Let $\phi:\mathbb{R}_+\to\mathbb{R}$ be a convex function with $\phi(1)=0$. If $P\ll Q$, then f-divergence is defined as $\mathrm{D}_\phi(P||Q)\triangleq\mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, where $\frac{dP}{dQ}$ is a Radon-Nikodym derivative.

The f-divergence family contains many popular divergences, such as KL divergence. Recently, [2] introduces a variational representation for f-divergence, as given below.

Lemma I.1. *Let ϕ^* be the convex conjugate of ϕ, and $\mathcal{G}=\{g:\Theta\to\mathrm{dom}(\phi^*)\}$. Then*

$$\mathrm{D}_\phi(P||Q)=\sup_{g\in\mathcal{G}}\mathbb{E}_{\theta\sim P}\left[g(\theta)\right]-\inf_{\alpha\in\mathbb{R}}\{\mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta)+\alpha)\right]-\alpha\}.$$

II. Main Results

Let $I^h_{\phi,\mu}(t\ell\circ h')=\inf_\alpha\mathbb{E}_\mu\left[\phi^*(t\ell(h,h')+\alpha)\right]-\alpha$. We now introduce a new discrepancy measure based on Lemma I.1.

Definition II.1 (f-DD)**.** For a given $h\in\mathcal{H}$,

$$\mathrm{D}^{h,\mathcal{H}}_\phi(\nu||\mu)\triangleq\sup_{h'\in\mathcal{H},t\in\mathbb{R}}\mathbb{E}_\nu\left[t\ell(h,h')\right]-I^h_{\phi,\mu}(t\ell\circ h').$$

We are in a position to give the target error bound.

Theorem II.1. *Let $\psi(x)\triangleq\phi(x+1)$, and ψ^* is its convex conjugate. Define $K_{h',\mu}(t)\triangleq\inf_\alpha\mathbb{E}_\mu\left[\psi^*(t\cdot\ell(h,h')+\alpha)\right]$. Let $K_\mu(t)=\sup_{h'\in\mathcal{H}}K_{h',\mu}(t)$. Then, for any $h\in\mathcal{H}$,*

$$R_\nu(h)\leq R_\mu(h)+\inf_{t\geq0}\frac{\mathrm{D}^{h,\mathcal{H}}_\phi(\nu||\mu)+K_\mu(t)}{t}+\lambda^*,\qquad(1)$$

where $\lambda^*=\min_{h^*\in\mathcal{H}}R_\mu(h^*)+R_\nu(h^*)$.

Given additional information about ϕ, $K_\mu(t)$ can be further upper bounded using a more expressive form, allowing for the determination of the optimal t. For instance, considering the KL case (denoted as $\mathrm{D}^{h,\mathcal{H}}_{\mathrm{KL}}$), the second term in Eq. (1) can be upper bounded by $\sqrt{2\mathrm{D}^{h,\mathcal{H}}_{\mathrm{KL}}(\nu||\mu)}$. Consequently, Theorem II.1 can recover previous KL-based results in [3].

TABLE I
Accuracy (%) on UDA Classification Tasks

Method	Office-31	Office-Home	Digits
[1]	89.5	68.5	96.3
Ours	**90.1**	**70.2**	**97.1**

Theorem II.1 suggests that by jointly minimizing the error of the source domain and the f-DD between two domains, a reduction in target error can be achieved. As such, we integrate a UDA algorithm similar to that proposed in [1], and our algorithm outperforms [1] as presented in Table I.

References

[1] D. Acuna, G. Zhang, M. T. Law, and S. Fidler, "f-domain adversarial learning: Theory and algorithms," in *International Conference on Machine Learning*. PMLR, 2021, pp. 66–75.
[2] R. Agrawal and T. Horel, "Optimal bounds between f-divergences and integral probability metrics," in *International Conference on Machine Learning*. PMLR, 2020, pp. 115–124.
[3] Z. Wang and Y. Mao, "Information-theoretic analysis of unsupervised domain adaptation," in *International Conference on Learning Representations*, 2023.