uai2025

Learning from Label Proportions and Covariate-shifted Instances

Sagalpreet Singh, Navodita Sharma, Shreyas Havaldar, Rishi Saket, Aravindan Raghuveer

DeepMind

Introduction

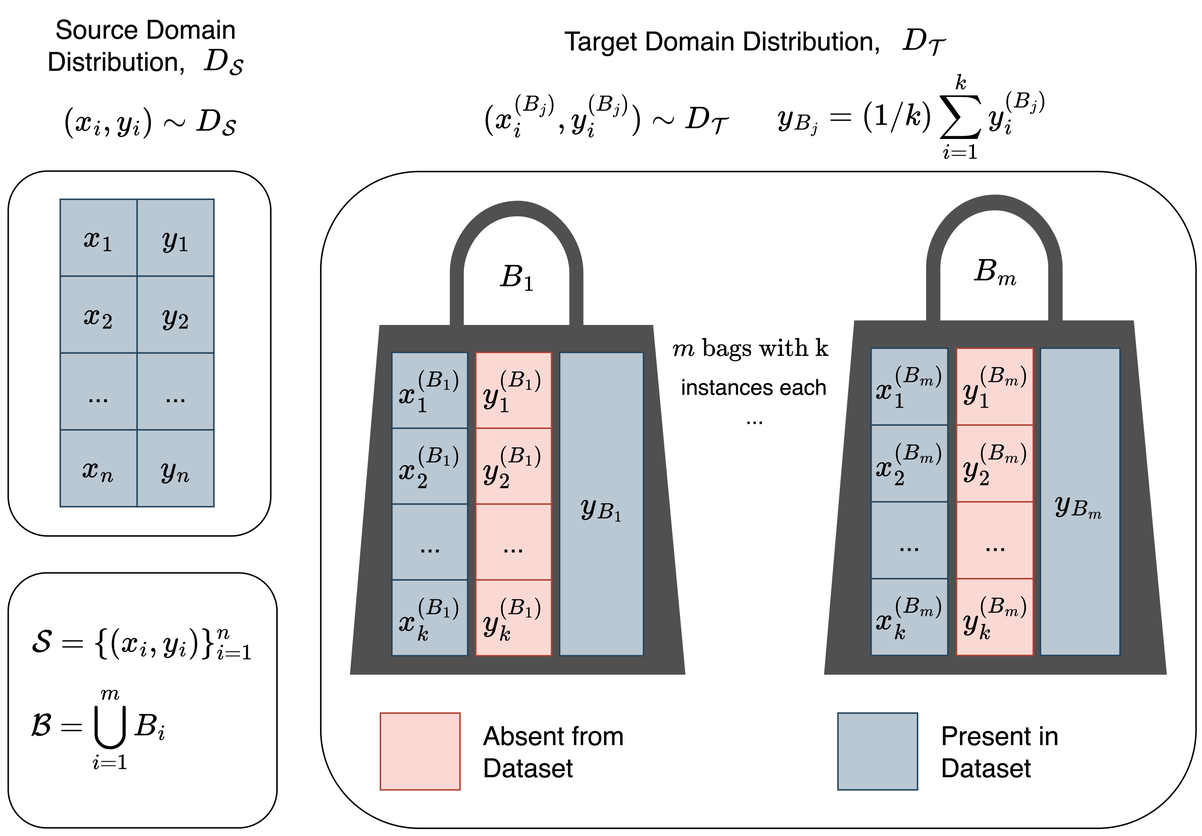
In many real-world scenarios, data is often organized into ”bags” where only an aggregate label for the entire bag is available. Additionally, we often have access to fully supervised source data that is covariate-shifted. Our work explores the use of valuable training signals from this source data in domain adaptation to improve instance-level prediction on the target domain.

Preliminaries. In traditional, fully supervised regression, the training data consists of labeled feature-vectors (instances) and corresponding labels. In *Learning from label proportions (LLP)*, feature-vectors are grouped into bags and only a *bag-label* is available for each bag which is the average of the instance-labels in the bag. In *Domain adaptation*, the challenge is to leverage data from a different but related *source* domain to perform well on a *target* domain. Domain adaptation under *covariate-shift* considers domains for which p(y|x) is the same but p(x) differs. We consider *covariate-shifted hybrid LLP* in which the source domain has instance-labels while the target has bag-labels.

Our key **contributions** are:

1. Novel methods within domain adaptation framework integrating both the target bag-labels and the source instance-labels.
2. Theoretical guarantees bounding the target generalization error.
3. Extensive experiments on various public datasets.

Problem Setting



- Source distribution, $D_{\mathcal{S}}$ is covariate-shifted from target distribution, $D_{\mathcal{T}}$.
- Dataset comprises instance labeled data points from source distribution and bag labeled data points from target distribution.
- Bags in the target domain are random partitions (size k each) of the target domain data.
- The goal is to learn a predictor to make instance level predictions on target domain on a regression task.

Notation

Consider a class $\mathcal{F}$ of real-valued functions (regressors). Denote the prediction function $h \in \mathcal{F}$ by $h(x) = r_h^T \varphi(x)$ where φ is a mapping to a real-vector in an embedding space and r_h is the representation of h in that space.

We define 3 different loss terms — distribution loss $\left(\varepsilon(\mathcal{D}, h) = \mathbb{E}_{(x,y)\leftarrow D}\left[(h(x)-y)^2\right]\right)$, finite sample loss $\left(\hat{\varepsilon}(\mathcal{U}, h) = \frac{1}{|\mathcal{U}|}\sum_{(x,y)\in U}\left[(h(x)-y)^2\right]\right)$ and loss on bags $\left(\bar{\varepsilon}(\mathcal{B}, h) = \frac{1}{|\mathcal{B}|}\sum_{(B,y_B)\in\mathcal{B}}\left(\left(\frac{1}{|B|}\sum_{x\in B} h(x)\right) - y_B\right)^2\right)$.

$\xi(\mathcal{S}, \mathcal{B})$ is the covariate-shift loss defined by us which takes into account the instance labels from source domain as well as bag labels from target domain. $\lambda'(\mathcal{S}, \mathcal{T})$ is a function independent of h. $R(h, \mathcal{S}, \mathcal{T})$ is a label-independent regularization on $\mathcal{S}$ and $\mathcal{T}$.

$$\xi(\mathcal{S}, \mathcal{B}) := 2\left\|\frac{1}{m}\sum_{j=1}^{m} y_{B_j}\left(\frac{1}{k}\sum_{x\in B_j}\varphi(x)\right) - \frac{1}{mk}\sum_{i=1}^{mk} l_i\varphi(z_i)\right\|_2$$

$$\lambda'(\mathcal{S}, \mathcal{T}) = \left|1/(mk)\sum_{i=1}^{mk}\left(y_i^2 - l_i^2\right)\right|, \quad R(h, \mathcal{S}, \mathcal{T}) = \left|1/(mk)\sum_{i=1}^{mk}\left(h(x_i)^2 - h(z_i)^2\right)\right|$$

Theoretical Results

Lemma
For any $h \in \mathcal{F}$,

$$\bar{\varepsilon}(\mathcal{B}, h) - \hat{\varepsilon}(\mathcal{S}, h) \le \xi(\mathcal{S}, \mathcal{B})\left\|r_h\right\|_2 + \lambda'(\mathcal{S}, \mathcal{T}) + R(h, \mathcal{S}, \mathcal{T})$$

Theorem
For $m, k \in^+, \nu, \delta > 0$, w.p. $1-\delta$ over choice of $\mathcal{B} = \mathcal{B}(m, k)$,

$$(\mathcal{D}_T, h) \le 16k\bar{\varepsilon}(\mathcal{B}, h)$$

for all $h \in \mathcal{F}$ s.t. $(\mathcal{D}_T, h) \ge \nu$ and $p = \text{Pdim}(\mathcal{F})$, when

$$m \ge O\left(\left(p\left(\log\left(\frac{k}{\nu}\right) + \log\log\left(\frac{1}{\delta}\right)\right) + \log\frac{1}{\delta}\right)\max\left\{\frac{1}{k\nu^2}, \frac{k^2}{\nu}\right\}\right)$$

Error bound in the above theorem weakens with increasing bag size. The above lemma can be used to mitigate this weakening. This leads to the following loss formulation:

$$\text{BagCSI}\left(\mathcal{S}, \mathcal{B}, h, \{\lambda_i\}_{i=1}^3\right) := \lambda_1\bar{\varepsilon}(\mathcal{B}, h) + \lambda_2\hat{\varepsilon}(\mathcal{S}, h) + \lambda_3\xi^2(\mathcal{S}, \mathcal{B}) \quad (1)$$

BL-WFA and PL-WFA

We propose and experiment with 2 different algorithms BL-WFA and PL-WFA. We also note that these two algorithms are a part of a larger class of algorithms.

- BL-WFA optimizes w.r.t. BagCSI loss.
- ξ is the L_2 distance between label weighted mean embeddings on source and target distributions.
- Instance level labels are not available in target domain. Psuedo labels are used as weights.
- Depending upon the choice of pseudo labeling technique used, different loss algorithms are obtained.

Denote the pseudo labeling function as f_L. The choice of pseudo labeling function for BL-WFA and PL-WFA are as follows:

$$f_L^{BL}(x, B, y_B, h) = y_B, \quad f_L^{PL}(x, B, y_B, h) = h(x) + \left(y_B - (1/|B|)\sum_{x\in B} h(x)\right)$$

Training Algorithm

Input: $\mathcal{S}$, $\mathcal{B}$, $\{\lambda_i\}_{i=1}^3$, f_L, a, opt **Output:** h

Initialize h s.t. $\theta_h = \{\varphi_h, r_h\}$

for *minibatch* $\mathcal{S}_i \in \mathcal{S}$, *minibatch* $\mathcal{T}_i \in \mathcal{T}$ **do**

$e_{\mathcal{S}_i} = \sum_{x,y\in\mathcal{S}_i} y.\varphi_h(x)$, $e_{\mathcal{T}_i} = \sum_{B,y_B\in\mathcal{T}_i}\sum_{x\in B} f_L(x, B, y_B, h).\varphi_h(x)$

$\mathcal{L}_{DA} \leftarrow \|e_{\mathcal{S}_i} - e_{\mathcal{T}_i}\|^2$

$\mathcal{L}_{\mathcal{S}} \leftarrow \sum_{x,y\in\mathcal{S}_i}(h(x)-y)^2$, $\mathcal{L}_{\mathcal{T}} \leftarrow \sum_{B,y_B\in\mathcal{T}_i}((1/|B|)\sum_{x\in B} h(x)) - y_B)^2$

$\mathcal{L} = \lambda_1\mathcal{L}_{\mathcal{S}} + \lambda_2\mathcal{L}_{\mathcal{T}} + \lambda_3\mathcal{L}_{DA}$

$\theta_h \leftarrow$ gradient_update(opt, $\theta_h, a, \mathcal{L}$)

return h

Depending upon choice of f_L, BL-WFA ($f_L = f_L^{BL}$) and PL-WFA ($f_L = f_L^{PL}$) are obtained.

Experiments & Results

Dataset	IPUMS US Census				Wine			
Method \ Bag Size	8	32	128	256	8	32	128	256
Bagged-Target	1.14 ± 0.00	1.16 ± 0.00	1.22 ± 0.00	1.31 ± 0.01	**173.5 ± 0.4**	**177.7 ± 1.2**	191.0 ± 2.5	206.9 ± 3.5
AF	1.23 ± 0.01	1.31 ± 0.01	1.41 ± 0.02	1.43 ± 0.02	186.8 ± 2.1	190.3 ± 2.8	191.0 ± 2.4	192.4 ± 1.8
LR	1.15 ± 0.00	1.18 ± 0.00	1.24 ± 0.01	1.29 ± 0.01	185.9 ± 2.0	191.6 ± 1.6	193.8 ± 0.8	194.5 ± 1.0
AF-DANN	1.25 ± 0.02	1.33 ± 0.07	1.39 ± 0.07	1.39 ± 0.02	187.6 ± 1.7	190.5 ± 1.7	191.2 ± 2.5	191.9 ± 2.1
LR-DANN	1.16 ± 0.00	1.23 ± 0.02	1.51 ± 0.07	1.61 ± 0.13	186.2 ± 1.5	192.1 ± 2.0	193.7 ± 2.4	193.8 ± 2.5
DMFA	1.15 ± 0.00	1.18 ± 0.00	1.26 ± 0.01	1.30 ± 0.01	186.1 ± 1.7	191.8 ± 2.1	193.5 ± 2.4	194.5 ± 0.9
PL-WFA (our)	1.15 ± 0.00	1.18 ± 0.00	1.25 ± 0.01	1.29 ± 0.01	183.0 ± 0.6	186.6 ± 1.0	189.0 ± 0.8	188.9 ± 1.2
BL-WFA (our)	**1.14 ± 0.00**	**1.16 ± 0.00**	**1.22 ± 0.00**	**1.25 ± 0.01**	180.9 ± 0.5	184.6 ± 0.7	**186.0 ± 0.8**	**186.4 ± 0.5**
Dataset	**Synthetic**				**Criteo SSCL**			
Method \ Bag Size	**8**	**32**	**128**	**256**	**64**	**128**	**256**	**512**
Bagged-Target	0.71 ± 0.05	5.49 ± 0.93	17.87 ± 0.49	19.95 ± 0.34	208.78 ± 2.7	234.32 ± 3.3	254.78 ± 5.3	264.74 ± 5.3
AF	0.96 ± 0.07	6.22 ± 0.81	18.16 ± 0.50	20.00 ± 0.86	297.95 ± 6.5	296.51 ± 6.1	294.86 ± 5.3	299.93 ± 6.5
LR	0.71 ± 0.04	5.15 ± 1.06	18.10 ± 0.40	19.92 ± 1.55	207.78 ± 2.7	232.72 ± 10.	256.68 ± 13.	264.46 ± 5.4
AF-DANN	1.23 ± 0.06	8.16 ± 0.54	18.04 ± 0.95	20.15 ± 0.49	296.95 ± 6.3	296.35 ± 6.4	295.49 ± 5.2	297.91 ± 7.3
LR-DANN	1.02 ± 0.04	7.84 ± 0.87	17.76 ± 0.24	19.72 ± 0.29	206.39 ± 2.3	230.84 ± 3.1	243.62 ± 4.5	265.33 ± 4.6
DMFA	**0.69 ± 0.05**	4.39 ± 0.84	16.50 ± 1.47	19.07 ± 1.16	207.60 ± 2.7	232.40 ± 9.9	247.66 ± 3.4	264.51 ± 5.5
PL-WFA (our)	0.75 ± 0.06	4.43 ± 0.81	15.60 ± 0.94	18.40 ± 0.74	204.71 ± 2.6	226.39 ± 2.9	240.55 ± 3.3	254.46 ± 5.5
BL-WFA (our)	0.75 ± 0.05	**2.22 ± 0.22**	**10.36 ± 3.15**	**13.76 ± 0.60**	**204.62 ± 2.4**	**226.33 ± 2.9**	**240.39 ± 3.2**	**254.36 ± 5.5**

MSE scores for different methods and bag sizes on different datasets. Lower is better. These results are based on uniformly-random same-sized bags. The paper also contains results with variable bag sizes and correlated bags.

[1] X. Li and A. Culotta, “Domain adaptation for learning from label proportions using domain-adversarial neural network,” *SN Comput. Sci.*, vol. 4, no. 5, p. 615, 2023.

[2] E. M. Ardehaly and A. Culotta, “Proc. IJCAI,” pp. 3670–3676, 2016.

[3] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *International conference on machine learning*, pp. 97–105, PMLR, 2015.

sagalpreet@google.com