Addressing Parameter Choice Issues in Unsupervised Domain Adaptation by Aggregation

selected as **notable-top-5% paper at ICLR 2023**

Marius-Constantin Dinu,[1,2] Markus Holzleitner,[1] **Maximilian Beck**,[1] Duc Hoan Nguyen,[3] Andrea Huber,[1]
Hamid Eghbal-zadeh,[1] Bernhard A. Moser,[4] Sergei V. Pereverzyev,[3] Sepp Hochreiter,[1] Werner Zellinger [3]

Introduction

- In Unsupervised Domain Adaptation we have labelled source data $\{(x_i, y_i)\}_{i=1}^s \sim p$ and unlabelled target data $\{x'_i\}_{i=1}^t \sim q_X$
- The goal is to learn a model $f : X \to Y \subset \mathbb{R}^d$ with small target error $\mathcal{E}_q(f) := \int_{X\times Y} \|f(x) - y\|_Y^2 \, \mathrm{d}q(x, y)$
- **Problem: How do we choose hyperparameters (e.g., learning rate or regularization parameters) without target labels?**

State of the Art

- Step 1: Compute different models $f_1, \ldots, f_m : X \to Y$ by running the learning algorithm with different hyperparameters.
- Step 2: Select the model $f^{\mathrm{sel}} := \underset{f\in\{f_1,\ldots,f_m\}}{\arg\min} \; \mathcal{E}_q(f)$ with smallest target error

TL;DR

- **Selecting hyperparameters without target labels is hard**
- **Current methods select the single best model**
- **We compute a linear aggregation of all models by importance weighted least squares**
- **We give target error guarantees for the linear aggregation**

Method: Importance Weighted Linear Aggregation by Least Squares (IWA)

- Idea: Compute linear aggregation $f^{\mathrm{agg}} := \sum_{i=1}^m c_i f_i$ such that the target error is minimized:
  $\mathcal{E}_q\left(f^{\mathrm{agg}}\right) = \min_{c_1,\ldots,c_m\in\mathbb{R}} \mathcal{E}_q\Big(\sum_{i=1}^m c_i f_i\Big)$
- With this approach the error is smaller than the best single model: $\mathcal{E}_q(f^{\mathrm{agg}}) \le \mathcal{E}_q(f^{\mathrm{sel}})$
- We use vector-valued linear least squares to compute the aggregation weights and importance weighting to take the covariate shift into account
- **Tool 1: Vector-Valued Least Squares**

$$c^{\mathrm{agg}} := G^{-1}g = \underset{(c_1,\ldots,c_m)\in\mathbb{R}^m}{\arg\min} \int_X \Big\| \sum_{i=1}^m c_i f_i(x) - f_q(x) \Big\|_Y^2 \, \mathrm{d}q_X(x)$$

with Bayes predictor, $f_q(x) = \int_Y y \, \mathrm{d}q(y|x)$ and Gram matrix $G = \left(\int_X \langle f_k(x), f_u(x)\rangle_Y \, \mathrm{d}q_X(x)\right)_{k,u=1}^m$

Not computable!

$$g = \left(\int_X \langle f_q(x), f_k(x)\rangle_Y \, \mathrm{d}q_X(x)\right)_{k=1}^m$$

- **Tool 2: Importance Weighting for Covariate Shift**

  Under covariate shift assumption $p(y|x) = q(y|x)$ and bounded density ratio $\beta(x) := \frac{\mathrm{d}q_X}{\mathrm{d}p_X}(x) \in [0, B]$ it holds

$$g = \left(\int_X \langle f_p(x), f_k(x)\rangle_Y \, \beta(x) \, \mathrm{d}p_X(x)\right)_{k=1}^m$$ [Shimodaira 2000, Kanamori et al. 2009]

- **Algorithm 1:** Importance Weighted Least Squares Linear Aggregation (IWA).

Input : Set $f_1, \ldots, f_m : \mathcal{X} \to \mathcal{Y}$ of models, s labeled source samples $(\mathbf{x}, \mathbf{y})$ and t unlabeled target samples $\mathbf{x}'$.

Output : Linear aggregation $\widetilde{f} = \sum_{k=1}^m \widetilde{c}_k f_k$ with weights $\widetilde{c} = (\widetilde{c}_1, \ldots, \widetilde{c}_m) \in \mathbb{R}^m$.

Step 1 *Use unlabeled samples* $\mathbf{x}$ *and* $\mathbf{x}'$ *to approximate density ratio* $\frac{\mathrm{d}q_X}{\mathrm{d}p_X}$ *by some function* $\beta(x)$ *using a classical algorithm, e.g. Sugiyama et al. (2012).*

Step 2 *Compute weight vector* $\widetilde{c} = \widetilde{G}^{-1}\widetilde{g}$ *with empirical Gram matrix* $\widetilde{G}$ *and vector* $\widetilde{g}$ *defined by*

$$\widetilde{G} = \left(\frac{1}{t}\sum_{i=1}^t \langle f_k(x'_i), f_u(x'_i)\rangle_{\mathcal{Y}}\right)_{k,u=1}^m \qquad \widetilde{g} = \left(\frac{1}{s}\sum_{i=1}^s \beta(x_i)\,\langle y_i, f_k(x_i)\rangle_{\mathcal{Y}}\right)_{k=1}^m.$$

Return : Linear aggregation $\widetilde{f} = \sum_{k=1}^m \widetilde{c}_k f_k$.

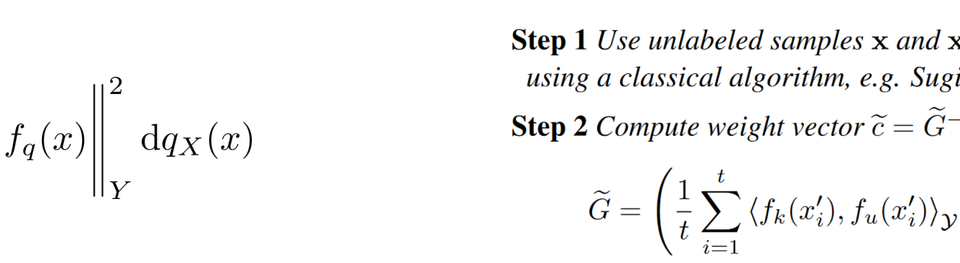
- **Theorem 1.** *With probability* $1-\delta$ *it holds that*

$$\mathcal{R}_q(\tilde{f}) - \mathcal{R}_q(f_q) \le 2\left(\mathcal{R}_q(f^*) - \mathcal{R}_q(f_q)\right) + C\left(\log\frac{1}{\delta}\right)(s^{-1} + t^{-1})$$

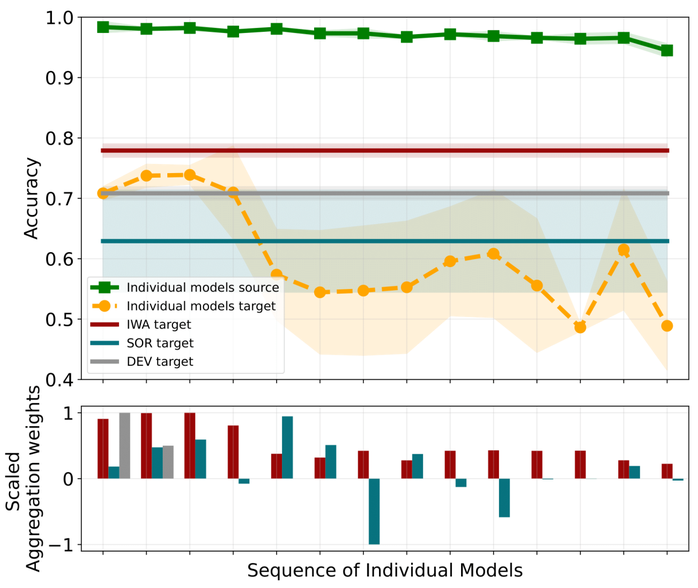
for some coefficient $C > 0$ *not depending on* s, t *and* δ*, and sufficiently large* s *and* t*.*

Results and Analysis

- We perform experiments on 6 benchmark datasets containing image, text and time series.
- For each dataset we use 12 domain adaptation methods and 6 parameter choice baselines (theoretically motivated and heuristic), which results in training about 16,000 models.
- We significantly outperform theoretically motivated model selection methods and beat each heuristic on at least 5 of 7 datsets
- In contrast to the other heuristic aggregation baselines, the aggregation weights of our method tend to be larger for accurate models



Histogram of the correlation coefficients of IWA and the linear regression heuristic baselines SOR, TCR and TMR over all datasets. IWA shows a stronger positive correlation between a model's target accuracy and its aggregation weight.



Top: Mean classification accuracy over 3 seeds. Bottom: Scaled aggregation weights for individual models. IWA effectively uses all models in the sequence.

Dataset	SO	TMV	TMR	TCR	SOR	IWV	DEV	IWA (ours)	TB
		Heuristic				Theoretical error guarantees			
Transformed Moons	0.989(±0.008)	0.980(±0.006)	0.981(±0.007)	*0.997(±0.002)*	0.989(±0.010)	0.989(±0.008)	0.981(±0.022)	**0.997(±0.002)**	0.997(±0.005)
Amazon Reviews	0.767(±0.011)	0.787(±0.009)	0.786(±0.010)	0.786(±0.010)	*0.789(±0.010)*	0.772(±0.014)	0.764(±0.019)	**0.788(±0.009)**	0.781(±0.012)
MiniDomainNet	0.507(±0.022)	*0.526(±0.011)*	0.525(±0.014)	0.526(±0.013)	0.518(±0.012)	0.513(±0.022)	0.515(±0.028)	**0.531(±0.011)**	0.534(±0.022)
Sleep-EDF	0.655(±0.054)	*0.729(±0.018)*	0.729(±0.024)	0.725(±0.023)	0.717(±0.028)	0.700(±0.052)	0.660(±0.057)	**0.737(±0.020)**	0.712(±0.045)
UCI-HAR	0.770(±0.046)	*0.840(±0.017)*	0.833(±0.023)	0.832(±0.024)	0.769(±0.060)	0.774(±0.070)	0.765(±0.090)	**0.835(±0.020)**	0.850(±0.029)
HHAR	0.732(±0.042)	0.771(±0.015)	0.768(±0.017)	*0.771(±0.018)*	0.722(±0.068)	0.746(±0.037)	0.722(±0.063)	**0.787(±0.012)**	0.784(±0.028)
WISDM	0.736(±0.050)	*0.768(±0.027)*	0.768(±0.036)	0.765(±0.037)	0.737(±0.062)	0.736(±0.052)	0.726(±0.077)	**0.764(±0.025)**	0.771(±0.046)

Avg. target accuracies and standard deviations over several domain adaptation tasks (e.g., 5 on HHAR, 5 on MiniDomainNet, 12 on Amazon Reviews), 11 domain adaptation methods (e.g., DANN, CMD, MMD) and 3 seeds.

1 ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning, Johannes Kepler University, Linz, Austria
2 Dynatrace Research
3 Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Linz, Austria
4 Software Competence Center Hagenberg

beck@ml.jku.at
maxmbeck

Paper: arxiv.org/abs/2305.01281

Video: iclr.cc/virtual/2023/oral/12613

Code: github.com/Xpitfire/iwa