Harnessing the Power of Vicinity-Informed Analysis for Classification under Covariate Shift

Mitsuhiro Fujikawa[1,3], Youhei Akimoto[1,3], Jun Sakuma[2,3], and Kazuto Fukuchi[1,3]

[1] University of Tsukuba [2] Institute of Science Tokyo [3] RIKEN AIP

Summary

Problem: we investigate classification problems under covariate shift:

- Input $X \in \mathcal{X}$, where $\mathcal{X}$ is a compact metric space equipped with a metric ρ and diameter $D_{\mathcal{X}}$.
- Label $Y \in \mathcal{Y}$, with $\mathcal{Y} = \{0, 1\}$.
- Source distribution P and target distribution Q, with a regression function $\eta : \mathcal{X} \to [0, 1]$ such that $P_{Y|X}(Y = 1|X) = Q_{Y|X}(Y = 1|X) = \eta(X)$ P_X- and Q_X-almost surely (covariate shift).
- Source sample $(\mathbf{X}, \mathbf{Y})_P = \{(X_i, Y_i)\}_{i=1}^{n_P} \sim P^{n_P}$ and target sample $(\mathbf{X}, \mathbf{Y})_Q = \{(X_i, Y_i)\}_{i=n_P+1}^{n_P+n_Q} \sim Q^{n_Q}$.

Goal: given $(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}, \mathbf{Y})_P \cup (\mathbf{X}, \mathbf{Y})_Q$, construct a classifier $h : \mathcal{X} \to \mathcal{Y}$ that minimizes

$$err_Q(h) = \mathbb{E}_Q \mathbb{1}\{h(X) \neq Y\}. \tag{1}$$

Analyses: we analyze the convergence rate of excess error for n_P and n_Q, defined as

$$\mathcal{E}_Q(h) = err_Q(h) - \inf_{h^*:\text{measurable}} err_Q(h^*). \tag{2}$$

Contributions:

- We construct an algorithm with source sample-size consistency, even under support non-containment conditions.
- Introduce Δ-transfer and Δ-self exponents to universally characterize convergence rate bounds of our and existing works, including Kpotufe et al. [2] and Pathak et al. [3], enabling fair comparision.
- Our convergence rate upper bound is always faster or competitive compared to Kpotufe et al. [2] and Pathak et al. [3].

Successful Transfer Learning and Source Sample-size Consistency

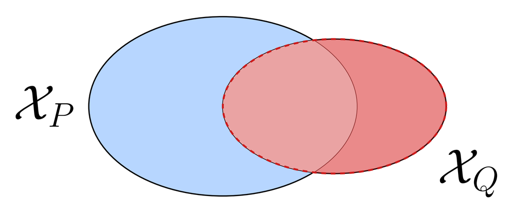
- A transfer learning algorithm is deemed successful if it achieves source sample-size consistency:

$$\sup_{P,Q} \mathbf{E}[\mathcal{E}_Q(h)] \to 0 \text{ as } n_Q \to \infty, \tag{3}$$

where the superemum is taken over an appropriate set of pairs of distributions.
- The source sample-size consistency indicates that the algorithm can reduce the error as the source sample-size increases.

Support Non-containment Environments

- Support of source distribution $\mathcal{X}_P = \{x \in \mathcal{X} : P_X(B(x, r)) > 0, \forall r > 0\}$.
- Support of target distribution $\mathcal{X}_Q = \{x \in \mathcal{X} : Q_X(B(x, r)) > 0, \forall r > 0\}$.
- Support non-containment : $\mathcal{X}_Q \not\subseteq \mathcal{X}_P$



Related Work

Properties of bounds under specific conditions.

	source sample-size consistency	support non-containment
Generalization error analyses		✓
Likelihood ratio-based	(✓)	
Likelihood ratio-based w/ support gap	(✓)	✓
[1–3]	✓	
our	✓	✓

Source sample-size consistency for likelihood ratio-based bounds requires access to the likelihood ratio function.

References

Nicholas R. Galbraith and Samory Kpotufe. Classification tree pruning under covariate shift. *IEEE Transactions on Information Theory*, 70(1):456–481, 2024. ISSN: 1557-9654. DOI: 10.1109/TIT.2023.3308914.

Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/21-AOS2084.

Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17517–17530. PMLR, 2022.

Check out arXiv version!

Dissimilarity measures

Pathak et al. [3]'s dissimilarity measure:

$$\Delta_{\mathrm{PMW}}(P, Q; r) = \int_{\mathcal{X}} \frac{1}{P_X(B(x, r))} Q_X(dx), \tag{4}$$

Δ_{PMW} becomes infinite under support non-containment environments.

Our dissimilarity measure:

$$\Delta_{\mathcal{V}}(P, Q; r) = \int_{\mathcal{X}} \inf_{x' \in \mathcal{V}(x)} \frac{1}{P_X(B(x', r))} Q_X(dx), \tag{5}$$

where

$$\mathcal{V}(x) = \left\{ x' \in \mathcal{X} : 2C_\alpha \rho(x, x')^\alpha < \left| \eta(x) - \frac{1}{2} \right| \right\} \cup \{x\}. \tag{6}$$

- $\mathcal{V}(x)$ denotes the set of the vicinity surrounding the point x.
- $\mathcal{V}(x)$ is the (nearly-)largest open ball centered at x with consistent labels.
- We may avoid zero division by taking the infimum over $\mathcal{V}(x)$.

Dissimilarity measure interpretation of Kpotufe et al. [2]:

$$\Delta_{\mathrm{DM}}(Q, Q; r) = \sup_{x \in \mathcal{X}_Q} \frac{1}{Q_X(B(x, r))}, \Delta_{\mathrm{BCN}}(Q, Q; r) = \mathcal{N}(\mathcal{X}_Q, \rho, r), \tag{7}$$

$$\Delta_{\mathrm{KM}}(P, Q; r) = \sup_{x \in \mathcal{X}_Q} \frac{Q_X(B(x, r))}{P_X(B(x, r))}. \tag{8}$$

Δ-transfer- and Δ-self-exponents

- (P, Q) has Δ-transfer-exponent τ if $\sup_{0<r\leq D_{\mathcal{X}}} (r/D_{\mathcal{X}})^\tau \Delta(P, Q; r) \leq C$.
- Q has Δ-self-exponent ψ if $\sup_{0<r\leq D_{\mathcal{X}}} (r/D_{\mathcal{X}})^\psi \Delta(Q, Q; r) \leq C$.

Given (P, Q), τ_Δ and ψ_Δ are the minimum exponents.

Main result

- Smoothness: $|\eta(x) - \eta(x')| \leq C_\alpha \cdot \rho(x, x')^\alpha$.
- Noise condition: $Q_X(0 < |\eta(X) - \frac{1}{2}| \leq t) \leq C_\beta t^\beta$.

	τ	ψ
Kpotufe et al. [2]	$\tau_{\Delta_{\mathrm{KM}}} + \min\{\psi_{\Delta_{\mathrm{DM}}}, \psi_{\Delta_{\mathrm{BCN}}}\}$	$\min\{\psi_{\Delta_{\mathrm{DM}}}, \psi_{\Delta_{\mathrm{BCN}}}\}$
Pathak et al. [3]	$\tau_{\Delta_{\mathrm{PMW}}}$	$\psi_{\Delta_{\mathrm{PMW}}}$
our	$\tau_{\Delta_{\mathcal{V}}}$	$\psi_{\Delta_{\mathcal{V}}}$

Universal rate

k-NN classifier w/ an appropriate k achieves

$$\mathbb{E}\left[\mathcal{E}_Q(\hat{h})\right] \leq C \begin{cases} \log(n_P + n_Q) \left(n_p^{\frac{1+\beta}{2+\beta+\max\{1,\tau/\alpha\}}} + n_Q^{\frac{1+\beta}{2+\beta+\max\{1,\psi/\alpha\}}} \right)^{-1} & \text{if } \alpha = \tau \text{ or } \alpha = \psi, \\ \left(n_p^{\frac{1+\beta}{2+\beta+\max\{1,\tau/\alpha\}}} + n_Q^{\frac{1+\beta}{2+\beta+\max\{1,\psi/\alpha\}}} \right)^{-1} & \text{otherwise .} \end{cases} \tag{9}$$

For any (P, Q),

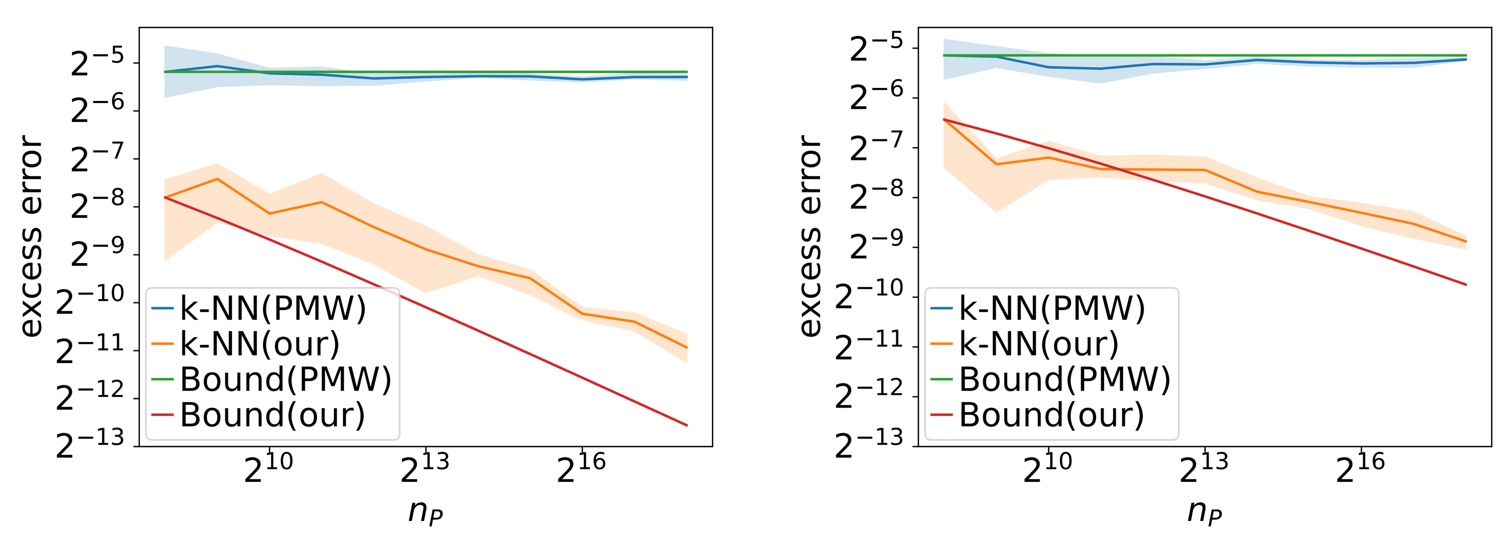
$$\tau_{\Delta_{\mathcal{V}}} \leq \tau_{\Delta_{\mathrm{PMW}}} \leq \tau_{\Delta_{\mathrm{KM}}} + \min\{\psi_{\Delta_{\mathrm{DM}}}, \psi_{\Delta_{\mathrm{BCN}}}\}, \tag{10}$$

$$\psi_{\Delta_{\mathcal{V}}} \leq \psi_{\Delta_{\mathrm{PMW}}} \leq \min\{\psi_{\Delta_{\mathrm{DM}}}, \psi_{\Delta_{\mathrm{BCN}}}\}, \tag{11}$$

- In the non-transfer setting, the exponent is $-\frac{1+\beta}{2+\beta+d/\alpha}$ for d-dimensional input.
- The expontns of our bound are equivalent to above, except d is replaced by the $\Delta_{\mathcal{V}}$-transfer- or $\Delta_{\mathcal{V}}$-self-exponent, corresponding to n_P or n_Q, respectively.
- $\Delta_{\mathcal{V}}$-self-exponent plays a role similar to the dimensionality d, as it is smaller than d.

Experiments

- $p_X(x) \propto (1 - x^2)^{-\tau/2}$, $\mathcal{X}_P = [-\frac{8^{\frac{1}{\alpha}} \cdot 2 - 1}{8^{\frac{1}{\alpha}} \cdot 2}, \frac{8^{\frac{1}{\alpha}} \cdot 2 - 1}{8^{\frac{1}{\alpha}} \cdot 2}]$, $\mathcal{X}_Q = [-1, 1]$, $\eta(x) = \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(x)|x|^\alpha$.



(a) $\alpha = \frac{1}{2}, \tau = 1$ (b) $\alpha = \frac{1}{2}, \tau = 2$