深層学習を用いたドメイン適応手法における速い収束レートについて

数理情報学専攻 48-186204　杉山 幹太
指導教員　鈴木 大慈 准教授

1　はじめに

学習に用いるデータの分布と予測を行いたいデータの分布が異なる場合を扱うドメイン適応と呼ばれる機械学習手法をドメイン適応手法という．特に予測を行いたいデータにラベルが無い場合を教師なしドメイン適応と呼び，本論文では簡単のため教師なしドメイン適応を単にドメイン適応と呼ぶこととする．学習に用いるデータの分布を元ドメイン, 予測を行いたいデータの分布を目標ドメインと呼ぶ.

ドメイン適応手法の中でも，深層学習を用いてドメイン適応を行う手法が近年多く研究されている．それはドメイン適応の汎化誤差の理論に基づいており，

$$R_T(h) \leq R_S(h) + \mathrm{Disc}(\mu_s, \mu_t) + \min_{h \in \mathcal{H}}(R_S(h) + R_T(h)) \tag{1}$$

が成立するという形式をしている [1]．ただし，$R_T(h)$ は仮説 h に対する目標ドメインの汎化誤差，$R_S(h)$ は元ドメインの汎化誤差，$\mathrm{Disc}(\mu_s, \mu_t)$ は元ドメインと目標ドメイン分布間の不一致度である．$\mathrm{Disc}(\mu_s, \mu_t)$ が小さくなるよう特徴抽出器を学習し，その特徴量で $R_S(h)$ が小さくなるよう学習するという手法が提案されている [2].

[4] はこれらの手法は一般には $R_T(h)$ を最小化しないという問題点を指摘しているが，この研究は特定の手法を用いた場合に関する指摘であり，別の手法によって問題を回避できる可能性を否定できないという問題がある．

また，式 (1) では任意の仮説 h に対して成り立つ式になっており，ドメイン適応の学習の結果得られた特定の仮説 $\hat{h}$ に対する評価式にはなっていない．そのため，第 3 項を無視したとしても $R_T(h)$ が収束するかどうかを示すことができないという問題がある．

本研究はこれらの問題を解決すべく，以下の 3 点を行う．

- 推定量の minimax リスクの下界を示すことで，どのような教師なしドメイン適応の手法もある問題設定があって誤差が定数だけ残ることを示す.
- 速い収束レートを実現する分布間不一致度 $\mathrm{Disc}(\mu_s, \mu_t)$ を提案する.
- 特徴抽出器 $g \in \mathcal{G}$ によって分布を動かすことを明示的に書き，実際にニューラルネットワークの集合 $\mathcal{F}_n, \mathcal{G}$ からドメイン適応手法による特徴抽出器 $\hat{g}$ と回帰のための関数 $\hat{f}$ という推定量を作り，適切な仮定なもと

$$R_T(\hat{f} \circ \hat{g}, h_t) - 16C_d \inf_{f^* \in \mathcal{F}_n}(R_T(f^* \circ \hat{g}, h_t) + R_S(f^* \circ \hat{g}, h_s))$$
$$= \tilde{\mathrm{O}}\left(n^{-\frac{2\beta}{2\beta+d}}\right)$$

という速い収束レートを示す.

2　ニューラルネットワークの定義

ニューラルネットワークとは, 整数 $L \geq 1, i = 1, ..., L$ として行列 $W_i \in \mathbb{R}^{p_{i-1} \times p_i}$，ベクトル $b_i \in \mathbb{R}^{p_i}$，活性化関数 σ に対し，$\sigma_i(\boldsymbol{x}) = \sigma(W_i x + b_i)$ と置くとき

$$f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}},\ f(\boldsymbol{x}) = W_{L+1}\sigma_L \circ ... \circ \sigma_1(\boldsymbol{x}),\ (2)$$

という形式で表される関数のことである．ニューラルネットワークをモデルとして機械学習を行うことを深層学習という．本研究では ReLU と呼ばれる活性化関数 $\sigma(x) = \max\{x, 0\}$ を用いたニューラルネットワークのみを考察する.

3　ノンパラメトリック回帰について

入力として i.i.d. に $X_1, ..., X_n \in [0,1]^d$ を観測，未知の関数 f_0 を用いて

$$Y_i = f_0(X_i) + \epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$$

としてデータ $(X_1, Y_1), ..., (X_n, Y_n)$ が得られる回帰問題を考える．本研究では，無限次元のモデルを対象の関数クラスとするノンパラリック回帰を考える．特に，半径 K の β-Hölder 空間

$$C_d^\beta(K)$$
$$= \left\{ f : [0,1]^d \to \mathbb{R} \;\middle|\; \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\beta} \|\partial^{\boldsymbol{\alpha}} f\|_\infty \right.$$
$$\left. + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\boldsymbol{x},\boldsymbol{y}\in[0,1]^d, \boldsymbol{x}\neq\boldsymbol{y}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|_\infty^{\beta - \lfloor\beta\rfloor}} \leq K \right\}$$

を推定対象の関数クラスとする．ただし，$\boldsymbol{\alpha}$ は多重指数である.

定理 1 ([5])**.** X_i の従う分布が $[0,1]^d$ 上のルベーグ測度に対して絶対連続な分布でその密度の下界と上界があ

る正の実数で抑えられているとする．するとある定数 $c > 0$ で

$$\min_{\tilde{f}:\text{推定量}} \max_{f_0 \in C_d^\beta(K)} R(\tilde{f}, f_0) \geq c n^{-\frac{2\beta}{2\beta+d}}$$

が成立する．

定理 1 の左辺を minimax レートと呼ぶ．定理 1 より，上界として $n^{-\frac{2\beta}{2\beta+d}}$ というレートが得られたならばそのレートは最適であることが分かる．

4 ドメイン適応における目標ドメインの汎化誤差の収束について

$y_i^s = h_s(x_i^s) + \epsilon_i$ とし，この h_s はある関数 $g_0 \in \mathcal{G}$ を用いて $h_s = f_s \circ g_0$ と分解でき，f_s は半径 K の β-Hölder 空間に属し，$g_{0\sharp}\mu_s = g_{0\sharp}\mu_t = \mu_g$ で，μ_g の台は $[0,1]^{d_g}$ に含まれるとする．この g_0 は正しい特徴抽出器と理解できる．$C_d > 1$ を任意の定数として，ドメイン間の不一致度を

$$\hat{d}_{\mathcal{F}_n}(g) = \sup_{f,f' \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^n (f \circ g(x_i^t) - f' \circ g(x_i^t))^2 - \frac{C_d}{n} \sum_{i=1}^n (f \circ g(x_i^s) - f' \circ g(x_i^s))^2 \right)$$

と定義する．また，元ドメインと目標ドメインのデータ数 n_s, n_t は $n_s = n_t =: n$ とする．

目標ドメインの汎化誤差について次の下界が成立する．

定理 2. $\mathcal{H}$ を可測関数 $h : \mathbb{R}^d \to \mathbb{R}$ からなる仮説集合とする．任意の $K > 0$ に対し

$$\inf_{\hat{h}} \sup_{f_t \in C_d^\beta(K)} R_T(\hat{h}, f_t \circ g_0) \geq \frac{K^2}{9}$$

が成立する．ただし inf は $(x_i^s, y_i^s, x_i^t)_{i=1}^n$ に関する，$\mathcal{H}$ に値を取る任意の推定量について取る．

目標ドメインの汎化誤差について次の上界が成立する．

定理 3. $\mathcal{G}$ は固定されたニューラルネットワークの関数集合とする．ニューラルネットワークの集合 $\mathcal{F}_n$ に対し $\hat{f}, \hat{g}$ を

$$\hat{f}, \hat{g} \in \operatorname*{arg\,min}_{f \in \mathcal{F}_n, g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n (f \circ g(x_i^s) - y_i^s)^2 + \hat{d}_{\mathcal{F}_n}(g) \right)$$

として

$$R_T(\hat{f} \circ \hat{g}, h_t) - 16 C_d \inf_{f^* \in \mathcal{F}_n} (R_T(f^* \circ \hat{g}, h_t) + R_S(f^* \circ \hat{g}, h_s)) = \tilde{\mathrm{O}}\left(n^{-\frac{2\beta}{2\beta+d}} \right)$$

を満たすあるニューラルネットワークの集合 $\mathcal{F}_n$ が存在する．

定理 3 で示した収束レートは通常の回帰の minimax レートと一致している．

5 考察

R_T に対する下界の定理 2 が成り立つのは，推定量が目標ドメインのラベル y_i^t を用いることができないためである．そのため，ラベルを付けることが可能な場合はラベル付けを行い教師なしドメイン適応ではなく（半）教師ありドメイン適応を行うべきである．

定理 3 により，学習の結果 $\inf_{f^* \in \mathcal{F}_n}(R_T(f^* \circ \hat{g}, h_t) + R_S(f^* \circ \hat{g}, h_s))$ という量が小さくなる場合は，$R_T(\hat{f} \circ \hat{g}, h_t)$ も小さくなることが言える．特に，

$$\inf_{f^* \in \mathcal{F}_n} (R_T(f^* \circ \hat{g}, h_t) + R_S(f^* \circ \hat{g}, h_s)) = \tilde{\mathrm{O}}\left(n^{-\frac{2\beta}{2\beta+d}} \right)$$

が成り立つとすると，定理 3 より

$$R_T(\hat{f} \circ \hat{g}, h_t) = \tilde{\mathrm{O}}\left(n^{-\frac{2\beta}{2\beta+d}} \right)$$

であり，目標ドメインにおける汎化誤差が速く収束することが言える．

正しい特徴抽出器を得られやすいと考えられる問題の例としては，$\mathcal{G}$ として平行移動を表す関数のみからなる集合を取る場合や，深層学習を用いたドメイン適応手法がある．

参考文献

[1] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 137–144, 2007.

[2] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[3] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

[4] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

[5] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.