2020年10月27日　統計数理研究所　オープンハウス

A General Class of Transfer Learning Regression without Implementation Cost

南　俊匠　　総合研究大学院大学　統計科学専攻　博士課程(5年一貫制)3年

始めに

- 転移学習(transfer learning)とは、元タスクで学習した知識を別のタスクに再利用する手法で、機械学習のフレームワークとして広く普及している。
- 特に、訓練データが限られており、ゼロからの学習が効果的ではない場合には、転移学習が対象タスクの予測性能を大幅に向上させる可能性があることが分かっている。
- 本研究の目的は、どのような回帰モデルにも適用可能な転移学習の新しいクラスを確立することである。
- 提案した手法は、いくつかの一般的な転移学習の手法を、統一された枠組みの中で扱うことができる。

問題設定

ソースタスク上の事前学習モデル$y = f_s(x)$が与えられた上で、ターゲット領域のn個のサンプルを用いて、$f_s(x)$をターゲットモデル$y = f_t(x)$に変換することが目的

既存手法の例

類似度に基づいた転移学習

$$L = \sum_{i=1}^{n}(y_i - f_{\theta_t}(x_i))^2 + \lambda||\theta_s - \theta_t||^2$$

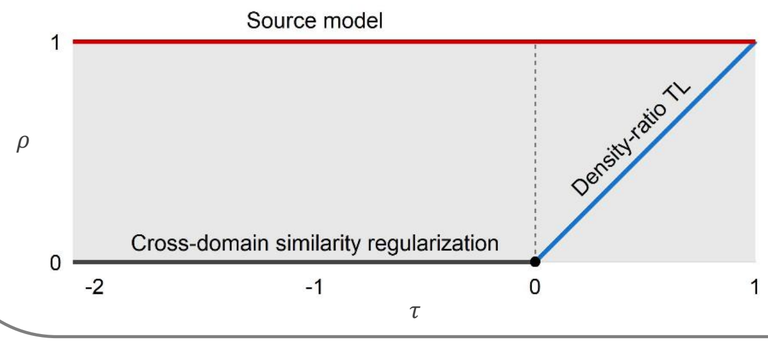
非類似度(差分)に基づいた転移学習

$$w(x) = f_t(x) - f_s(x) \quad \text{or} \quad w(x,y) = \frac{p_t(y|x)}{p_s(y|x)}$$

提案手法

目的関数　$$L = \sum_{i=1}^{n}\left[(y_i - f_{\theta_w}(x_i))^2 - \tau\left(f_s(x_i) - f_{\theta_w}(x_i)\right)^2\right]$$

予測　$$\hat{y}(x) = (1-\rho)f_{\hat{\theta}_w}(x) + \rho f_s(x) \qquad (\tau < 1,\ \ 0 < \rho < 1)$$



本手法がカバーする手法(左図)

- ソースモデルを直接使用
- 類似度に基づいた手法
- 非類似度に基づいた手法

特徴

特徴①　ハイパーパラメータにより転移のアプローチを選択することができる。

$\rho = 0, \tau < 0$: 類似度ベース

$$L = \sum_{i=1}^{n}\left[(y_i - f_{\theta_w}(x_i))^2 + \lambda\left(f_s(x_i) - f_{\theta_w}(x_i)\right)^2\right]$$

$$\hat{y}(x) = f_{\hat{\theta}_w}(x)$$

$\tau = \rho$: 非類似度ベース

$$L = \sum_{i=1}^{n}\left[(y_i - f_{\theta_w}(x_i))^2 - \rho\left(f_s(x_i) - f_{\theta_w}(x_i)\right)^2\right]$$

$$\hat{y}(x) = (1-\rho)f_{\hat{\theta}_w}(x) + \rho f_s(x)$$

特徴②　Bias-variance 分解

Bias+Variance

$$\mathrm{MSE}(\hat{y}(x)) = \left[\frac{\rho-\tau}{1-\tau}\mathrm{D}(x) + \frac{1-\rho}{1-\tau}\mathrm{B}_1(x) - \frac{\tau(1-\rho)}{1-\tau}\mathrm{B}_2(x)\right]^2 + \left(\frac{1-\rho}{1-\tau}\right)^2\mathrm{V}(x) + \sigma_\epsilon^2$$

where, $\mathrm{D}(x) = f_t(x) - f_s(x)$, $\mathrm{B}_1(x) = f_t(x) - x^{\mathrm{T}}\mathrm{S}\mathbf{f}_t$, $\mathrm{B}_2(x) = f_s(x) - x^{\mathrm{T}}\mathrm{S}\mathbf{f}_s$, $\mathrm{V}(x) = \sigma_\epsilon^2 x^{\mathrm{T}}\mathrm{S}\mathrm{S}^{\mathrm{T}}x$

特に、$\mathbb{E}[\mathrm{MSE}(\hat{y}(x))]$の最小化について

- $\mathbb{E}[\mathrm{D}^2]$が支配的であるとき、非類似度に基づく方法($\tau = \rho$)が選択される。
- $\mathbb{E}[\mathrm{V}] \to \infty$のときソースモデルを直接使用する方法($\rho = 1$)が選択される。

特徴③　実装コストがゼロである

目的関数は以下のように変形でき、作業変数zに対する回帰として実装することができる。

$$\hat{\theta}_w = \arg\min_{\theta_w}\sum_{i=1}^{n}(z_i - f_{\theta_w}(x_i))^2, \quad z_i = \frac{y_i - \tau f_s(x_i)}{1-\tau}.$$

実装手順

1. 訓練データの入出力の組 $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ を用意する。
2. データ (x_i, y_i) とソースのモデル $f_s(x)$ を使って、変数 z_i を作る。
3. 入力 x から変数 z を予測するモデル $f_{\hat{\theta}_w}(x)$ を訓練する。
4. 出力 y を以下の式で計算する。

$$\hat{y}(x) = (1-\rho)f_{\hat{\theta}_w}(x) + \rho f_s(x)$$

参考:複数タスクからの転移

フレームワーク

$$L = \sum_{i=1}^{n}\left[(y_i - f_{\theta_w}(x_i))^2 - \sum_k \tau_k\left(f_{s_k}(x_i) - f_{\theta_w}(x_i)\right)^2\right]$$

$$\hat{y}(x) = \rho_0 f_{\theta_w}(x) + \sum_k \rho_k f_{s_i}(x) \qquad \left(\sum_k \tau_k < 1,\ \ \sum_{k=0,1,\ldots}\rho_k = 1,\ \ 0 \le \rho_k \le 1\right)$$

Bias-variance 分解

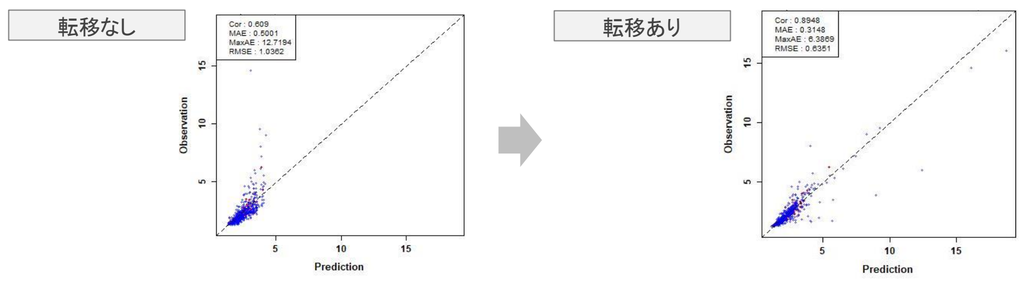
$$\mathrm{Bias} = \sum_k \frac{\rho_k - \tau_k}{1-\sum\tau_i}\mathrm{D}_{t,k}(x) + \frac{\rho_0}{1-\sum\tau_i}\mathrm{B}_0(x) - \sum_k \frac{\rho_0\tau_k}{1-\sum\tau_i}\mathrm{B}_k(x) + \sum_{k,l}\frac{\rho_k\tau_l}{1-\sum\tau_i}\mathrm{D}_{k,l}(x)$$

$$\mathrm{Var} = \left(\frac{\rho_0}{1-\sum\tau_i}\right)^2\mathrm{Var}(x)$$

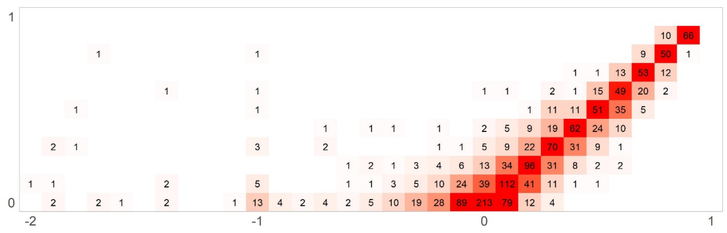
where, $\mathrm{D}_{t,k}(x) = f_t(x) - f_{s_k}(x)$, $\mathrm{D}_{k,l}(x) = f_{s_k}(x) - f_{s_l}(x)$, $\mathrm{B}_0(x) = f_t(x) - x^{\mathrm{T}}\mathrm{S}\mathbf{f}_t$, $\mathrm{B}_k(x) = f_{s_k}(x) - x^{\mathrm{T}}\mathrm{S}\mathbf{f}_{s_k}$

実験

実験①:　無機物質の誘電率の予測タスクを無機物質の屈折率の予測タスクへ転移



実験②:　555個のタスクについて、提案手法で選択されたハイパーパラメータの分布



まとめ・今後の展望

- 学習と予測におけるソースタスクの影響をコントロールする2つのハイパーパラメータを特徴とした新しいクラスの転移学習のフレームワークを提案した。
- ハイパーパラメータとモデルの選択に応じて、いくつかの手法がハイブリッド化された様々な学習方法を導出することができる。
- 実験的には、一般的な考え方である類似度に基づくアプローチだけでなく、それとは対立的な非類似度に基づくアプローチも多く選択された。
- 広範囲の損失関数や学習タスクを考慮した、より一般的な問題への拡張が今後の課題の一つである。

参考文献:　Minami et al. A general class of transfer learning regression without implementation cost. ArXiv (2020) arXiv:2006.13228