変数の拡張に対する最適輸送を用いたドメイン適応

有竹　俊光　　統計的機械学習研究センター　特任助教

概要

教師付き学習において教師，テストデータの従う分布が異なる場合，予測精度の低下をまねく．このような精度低下を解決する方法としてドメイン適応と呼ばれる方法が研究されている．本研究では，教師，テストデータが共通の変数を持ち，かつ，テスト時にのみ予測のための新しい変数が観測される場合を考える．そして，新規変数を用いたより高精度な推定を可能とするために，最適輸送を用いたドメイン適応法を提案する．

最適輸送を用いた教師なしドメイン適応

教師データ$\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{N_s}$とテストデータ$\{\boldsymbol{x}_j^t\}_{j=1}^{N_t}$の説明変数分布を近づけ，教師ラベルをテスト標本で利用

$$\hat{\pi}^* = \arg\min_{f\in\mathcal{F}, \hat{\pi}\in\hat{\Pi}(\mathcal{D}_S, \mathcal{D}_T)} \sum_{i,j} \hat{\pi}_{ij}\mathcal{E}_{\alpha,ij} \quad \textbf{(1)}$$

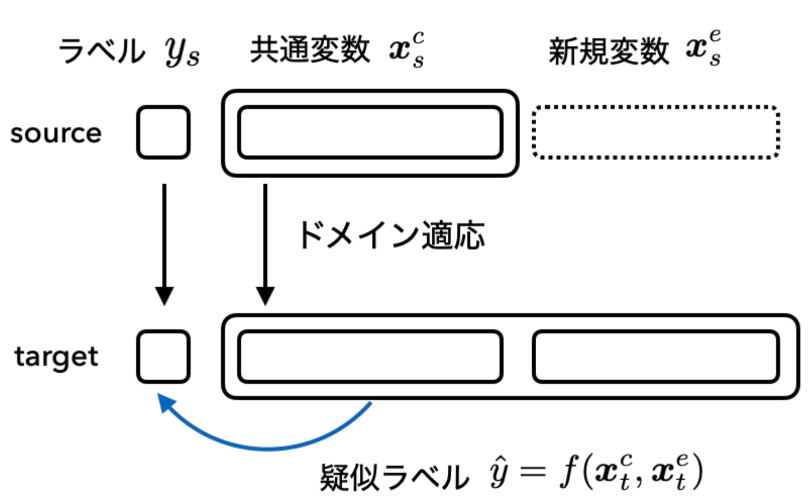
$$\hat{\Pi}(\mathcal{D}_S, \mathcal{D}_T) \equiv \left\{\pi \in \mathbb{R}_+^{N_s\times N_t} \middle| \sum_{i=1}^{N_s}\pi_{ij} = \frac{1}{N_t}, \sum_{j=1}^{N_t}\pi_{ij} = \frac{1}{N_s}\right\}$$

1. 輸送コストと輸送量の積の総和を最小化
   - 輸送コストとして変数間の距離とラベルの不一致度を利用 (Joint Distribution Optimal Transport; [1])
     $\mathcal{E}_{\alpha,ij} = \alpha\|\boldsymbol{x}_{si}^c - \boldsymbol{x}_{tj}^c\|_2^2 + \mathcal{L}(y_{si}, \hat{y}_{tj} = f(\boldsymbol{x}_{tj}^c))$
   - ターゲットラベルは未知なので疑似ラベル $\hat{y} = f(\boldsymbol{x})$ を利用
2. 教師ラベルを輸送先で利用してモデルを更新

変数の拡張に対するドメイン適応

問題

- テストデータでのみ新しく観測される変数が存在
  - 例：新しいセンサを追加してデータを計測して予測
  - 既存のセンサは同一の情報を計測しているが分布は異なる
  - 新規変数は教師データに含まれないため直接学習には利用不可能



- 既存変数の分布の違い，新規変数による次元の違いに対するドメイン適応の方法を提案

アルゴリズム

- 教師，テストデータの共通変数間の距離をコストとする
- 疑似ラベルで新規変数$\boldsymbol{x}^e$を考慮する $\hat{y}_{tj} = f(\boldsymbol{x}_{tj}^c, \boldsymbol{x}_{tj}^e)$

以下のステップを収束まで交互に反復

1. 式(1)と同様のJDOTにより教師データのラベルを輸送
   - 最初の輸送は共通変数間距離のみがコスト $\mathcal{E}_{\alpha,ij} = \alpha\|\boldsymbol{x}_{si}^c - \boldsymbol{x}_{tj}^c\|_2^2$
   - それ以降は $\mathcal{E}_{\alpha,ij} = \alpha\|\boldsymbol{x}_{si}^c - \boldsymbol{x}_{tj}^c\|_2^2 + \mathcal{L}(y_{si}, f^{(n)}(\boldsymbol{x}_{tj}^c, \boldsymbol{x}_{tj}^e))$
2. 輸送したラベルによりモデル f を学習
   $f^{(n+1)} = \mathrm{argmin}_{f\in\mathcal{F}} \sum_j \mathcal{L}(y_{si}, f^{(n)}(\boldsymbol{x}_{tj}^c, \boldsymbol{x}_{tj}^e))$

提案法の解釈

提案法は教師，テスト分布間の双方向の最適輸送と等価

1. ターゲット→ソースへの最適輸送により新規変数を推定
   $\mathcal{E}_{\alpha,ij} = \alpha\|\boldsymbol{x}_{si}^c - \boldsymbol{x}_{tj}^c\|_2^2 + \mathcal{L}(y_{si}, f(\boldsymbol{x}_{tj}^c, \boldsymbol{x}_{tj}^e))$
2. ソース→ターゲットへのラベルを輸送
   $\mathcal{E}_{\alpha,ij} = \alpha(\|\boldsymbol{x}_{si}^c - \boldsymbol{x}_{tj}^c\|_2^2 + \|\hat{\boldsymbol{x}}_{si}^e - \boldsymbol{x}_{tj}^e\|_2^2) + \mathcal{L}(y_{si}, f(\boldsymbol{x}_{tj}^c, \boldsymbol{x}_{tj}^e))$

- $\mathcal{P}_S(\boldsymbol{x}^e|\boldsymbol{x}^c, y) = \mathcal{P}_T(\boldsymbol{x}^e|\mathcal{T}(\boldsymbol{x}^c, y))$ を仮定
  - 輸送前後で共通変数のクラスごとの新規変数分布が同一
  - 新規変数の輸送コストが0となる輸送先が存在

理論解析

$$\mathrm{err}_T(y, \hat{f}_s) \le \underbrace{\boxed{\widehat{\mathrm{err}}_{\hat{T}}(\hat{f}_s, y) + \frac{2ka\Lambda}{\sqrt{N_t}} + (L_0 + ka\Lambda)\sqrt{\frac{\ln(1/\delta)}{N_t}}}}_{\text{ターゲットの近似分布での学習による汎化誤差の上界（Rademacher複雑度）}}$$

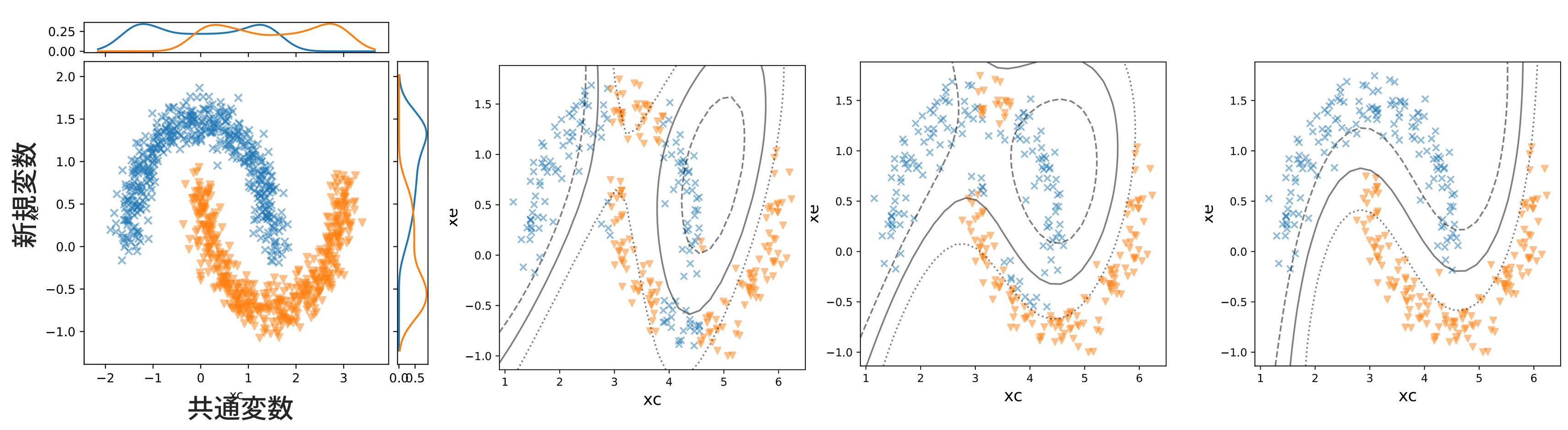
$$+\underbrace{W(\mathcal{P}_{\hat{T}}, \mathcal{P}_T)}_{\text{真のターゲット分布と近似分布のWasserstein距離}} + \underbrace{2\,\mathrm{err}_T(y, f_0) + kM\phi(\lambda)}_{\text{真のターゲット分布によって決まる定数項}},$$

- $\widehat{\mathrm{err}}_{\hat{T}}(\hat{f}_s, y)$ はテストデータの真の分布を近似する分布による経験誤差で $\hat{f}_s$はこの誤差を最小化
- ソース，ターゲット分布は次元が異なるためターゲットの近似分布とのWasserstein距離を考えている
- モデル集合 $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \le a\}$
  $\mathcal{H}$ は有界なカーネル K を備えた再生核ヒルベルト空間 $\sup_{\boldsymbol{x}} K(\boldsymbol{x}, \boldsymbol{x}) \le \Lambda^2$
- 誤差関数$\mathcal{L}$は以下を満たす
  - 対称 $\mathcal{L}(y_1, y_2) = \mathcal{L}(y_2, y_1)$，有界 $L_0 = \sup_{y\in\mathcal{C}} \mathcal{L}(0, y) < \infty$
  - 三角不等式を満たす $\mathcal{L}(y_1, y_2) + \mathcal{L}(y_2, y_3) \ge \mathcal{L}(y_1, y_3)$
  - k-リプシッツ連続 $|\mathcal{L}(y_1, y_2) - \mathcal{L}(y_1, y_3)| \le k|y_2 - y_3|$

実験

人工データ実験

- 新規変数として判別に有用な変数が観測されると仮定し提案法により学習された判別境界を可視化
- 教師データでは新規変数が観測されないので高精度な判別は不可能だが逐次的によい判別境界が得られる



実データ実験

- Gas Sensor Array Drift Dataset[2] を用いた実験
  - 各ドメインでセンサの劣化度合いが異なる

domains	Baseline	JDOT no extra	CCA	DSFT	Proposed	JDOT ideal
1→ 2	**83.33**	77.71	66.87	42.97	78.31	83.73
1→ 3	52.28	93.45	43.27	57.78	**96.02**	94.15
1→ 4	64.49	60.75	58.88	**94.39**	87.85	85.98
2→ 1	52.13	79.26	56.91	62.77	**84.04**	85.64
2→ 3	56.84	89.36	25.03	84.56	**89.47**	90.99
2→ 4	63.55	69.16	51.40	41.12	**71.96**	74.77
3→ 1	51.06	92.02	92.55	66.49	**94.15**	95.74
3→ 2	68.67	81.92	67.67	86.94	**88.76**	88.55
3→ 4	94.39	77.57	**96.26**	40.19	81.31	80.37
4→ 1	50.00	52.66	48.93	51.60	**53.72**	80.85
4→ 2	42.97	71.08	32.93	32.93	**74.30**	73.89
4→ 3	**92.98**	82.81	57.31	42.69	82.57	82.57

  - エタノール，エチレンの2種類のガスの判別精度を適応を行わない場合（Baseline）新規変数を考慮しない最適輸送（JDOT no extra），先行研究[3, 4]と比較
  - JDOT idealはソースでも新規変数が観測される理想的な場合のベンチマーク

まとめ

- 最適輸送を用いてテストデータで変数が増える場合に対するドメイン適応方法を提案
- 提案法に双方向輸送という解釈を与え，Rademacher複雑度，Wasserstein距離によるテスト誤差の上界を導出
- 本研究内容はWCCI2022[5]で発表予定

参考文献

[1] Courty, et al., “Joint distribution optimal transportation for domain adaptation”, Adv. Neural Inf. Process. Syst. (2017)
[2] Vergara, et al., “Chemical gas sensor drift compensation using classifier ensembles”, Sensors and Actuators B: Chemical (2012)
[3] Yeh, et al., “Heterogeneous domain adaptation and classification by exploiting the correlation subspace”, IEEE Trans. Image Process. (2014)
[4] Wei, et al., “A General Domain Specific Feature Transfer Framework for Hybrid Domain Adaptation”, IEEE Trans Knowl Data Eng., (2019)
[5] Aritake, et al., “Domain Adaptation with Optimal Transport for Extended Variable Space”, IEEE World Congress on Computational Intelligence. (2022), to be appear.