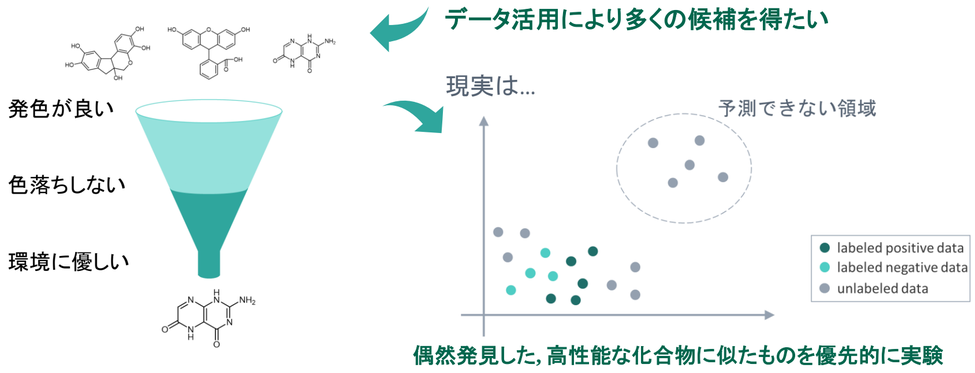
Multifidelity能動学習を用いた段階的ドメイン適応

佐川　正悟　　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

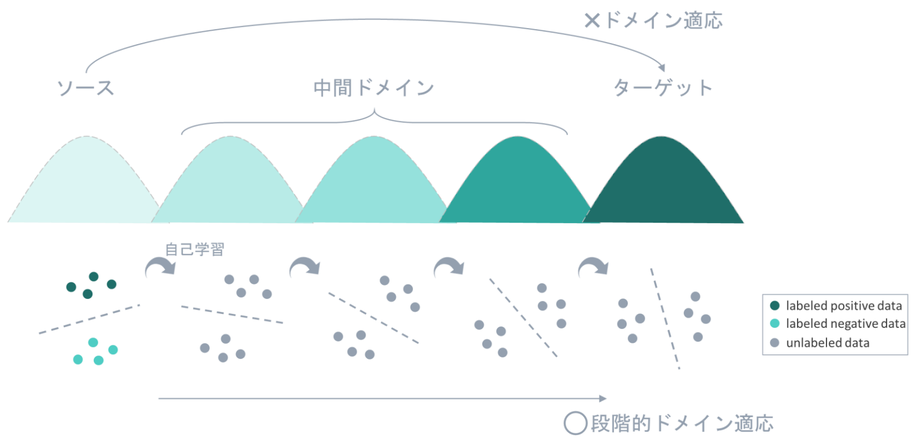
【背景 | 材料開発におけるデータ活用の課題】

材料開発は顧客要望の高いものから技術開発を行い, 複数の機能を満たすものが最後に選択される.
要素技術開発の段階で, 化学構造の多様性が高ければ, 製品化の確率が高くなる.



偶然発見した, 高性能な化合物に似たものを優先的に実験

【先行研究 | 段階的ドメイン適応】

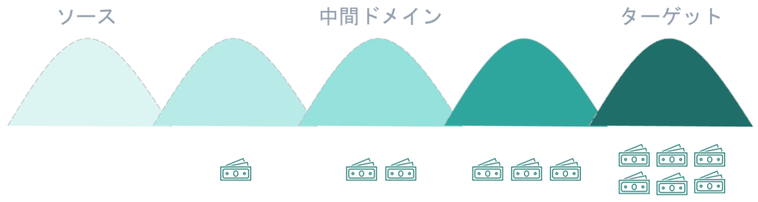


Kumar, Ananya, Tengyu Ma, and Percy Liang. "Understanding self-training for gradual domain adaptation." *International Conference on Machine Learning*. PMLR, 2020.

【研究目的, 問題設定】

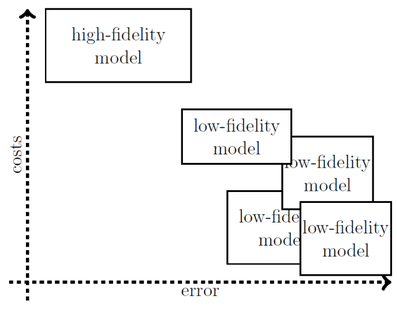
◎研究目的
利用可能な中間ドメインが限られている場合でも, 段階的ドメイン適応を可能とする！

利用可能な中間ドメインが限られている状況を解決するために, 少量のクエリを要求することにする.
ただし, 各ドメインにはクエリコストが設定されていることにする.
ターゲットのデータのみで, モデルを作ろうとするとコスト↑.



●クエリコスト設定について
ターゲットにラベルが無い理由をコストが高いためと想定.
Ex. 材料開発: ソースは材料が安い, ターゲットは材料が高い.
　　病理画像診断: ソースは医師でなくても判断が容易, ターゲットは専門の医師でないと判別困難.

【関連研究 | Multifidelity学習】



※下記、参考文献より引用

Peherstorfer, Benjamin, Karen Willcox, and Max Gunzburger. "Survey of multifidelity methods in uncertainty propagation, inference, and optimization." *Siam Review* 60.3 (2018): 550-591.

【提案手法 | 学習モデル】

段階的ドメイン適応では, 隣接ドメイン間の潜在変数やラベル情報は類似する可能性が高い.
段階的ドメイン適応に適した学習モデルとしてGDAMF (Gradual Domain Adaptation with Multifidelity) を考案.

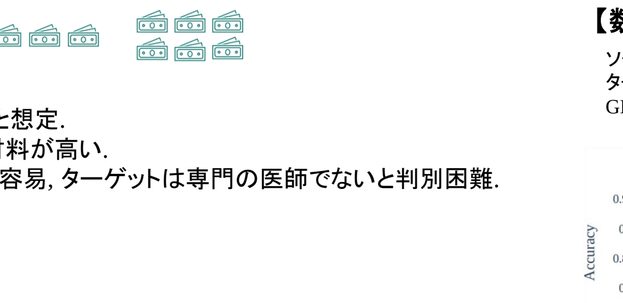
特徴量抽出器

ソース用

$$f_\alpha : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}, \quad \boldsymbol{x} \mapsto \boldsymbol{z}^{(0)}$$

中間ドメイン/ターゲット用

$$\varphi_{\gamma_q} : \mathbb{R}^{d_2} \to \mathbb{R}^{d_2}, \quad \boldsymbol{z}^{(q-1)} \mapsto \boldsymbol{z}^{(q)}$$



分類器

ソース用

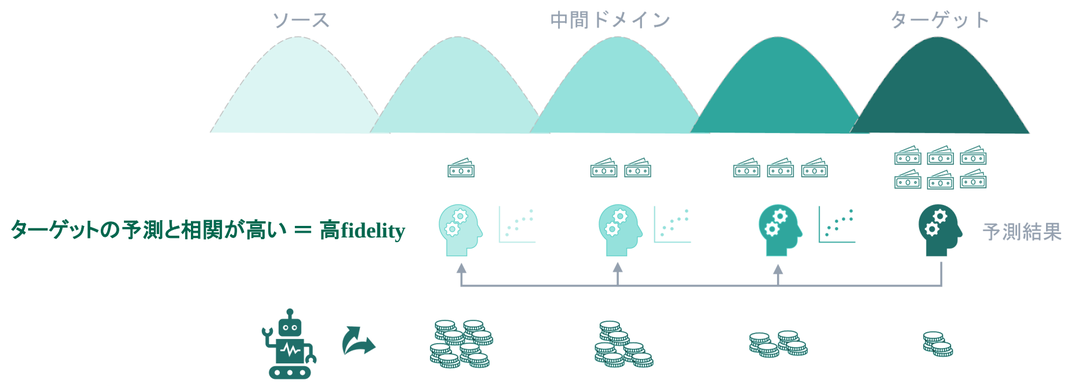
$$g_\beta : \mathbb{R}^{d_2} \to \Delta_N, \quad \boldsymbol{z}^{(0)} \mapsto (p_1, p_2, \ldots, p_N)$$

中間ドメイン/ターゲット用

$$\psi_{\theta_q} : \mathbb{R}^{d_2} \times \Delta_N \to \Delta_N, \quad (\boldsymbol{z}^{(q)}, M'_{q-1}(\boldsymbol{x})) \mapsto (p_1, p_2, \ldots, p_N)$$

【提案手法 | Multifidelity能動学習】

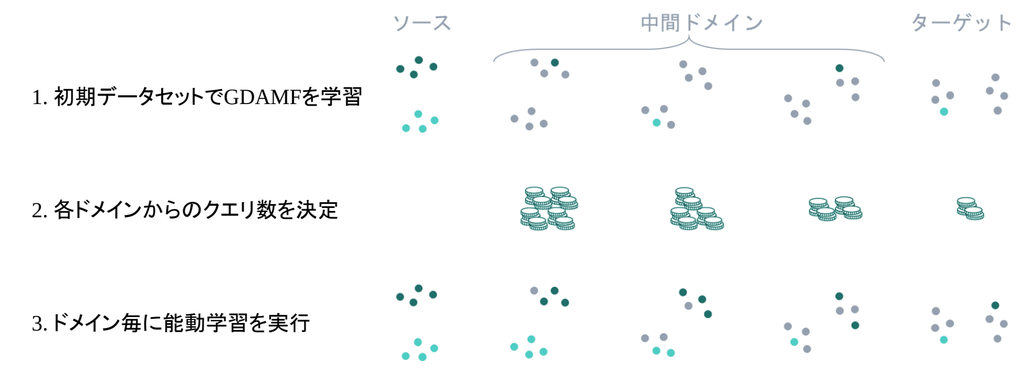
Multifidelityモンテカルロ法を応用し, コストとfidelityを考慮して各ドメインからのクエリ数を決定する.
決められたクエリ数分, ドメイン毎に能動学習を実行し, モデリングに有用なデータを集める.



高fidelityかつ低コストなドメインからのクエリを優先し, 予算を分配する

Peherstorfer, Benjamin, Karen Willcox, and Max Gunzburger. "Optimal model management for multifidelity Monte Carlo estimation." *SIAM Journal on Scientific Computing* 38.5 (2016): A3163-A3194.

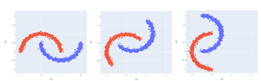
【提案手法 | 全体の流れ】

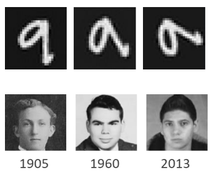


【データセット】

先行研究に倣い, 同様のデータセットを用意し, 使用できる中間ドメインに制限をかける.

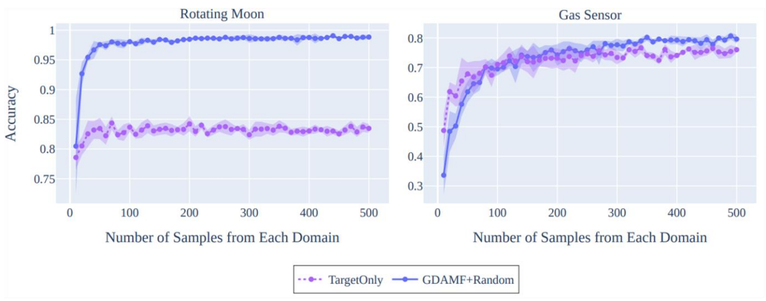
- Rotating Moon
  two-moonをソースとし, 回転を加えたものを中間ドメイン (3) /ターゲットとする.
- Rotating MNIST
  回転を加えたものを中間ドメイン (21) /ターゲットとする.
- Portraits
  画像から男女を当てる. 年代順に古い方から, ソース/中間ドメイン (14) /ターゲットとする.
- Cover type
  54の特徴量から植物の種類を当てる. 水辺からの距離が遠い方から, ソース/中間ドメイン (30) /ターゲットとする.
- Gas sensor
  128の特徴量からガスの種類を当てる. センサーの劣化によりシフトが起きる.
  劣化していない方からソース/中間ドメイン (7) /ターゲットとする.





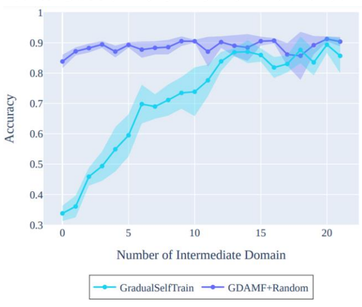
【数値実験】

ソースに加え, 中間ドメインのデータを活用することの有効性を確認する.
ターゲットのデータのみで構築したモデルと, GDAMFを比較する.
GDAMFの場合は, 全ての中間ドメインとターゲットから指定された数をランダムサンプリング.

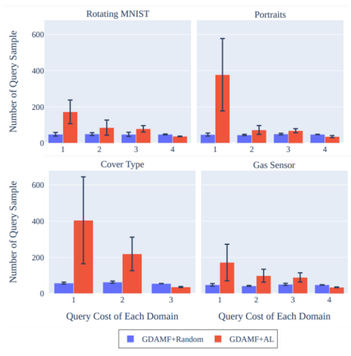


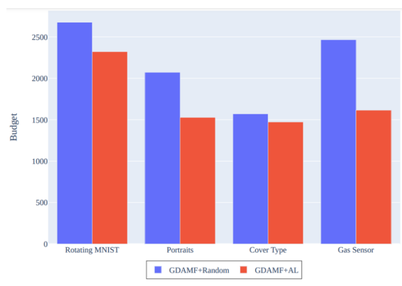
ソース/中間ドメイン/ターゲットのデータを活用するGDAMFの方が高性能

既存手法 (自己学習による段階的ドメイン適応) との比較を示す.
Rotating MNISTにて全部で21ある中間ドメインの内, 使用できるドメインの数を変化させる. GDAMFの場合は, 利用可能な中間ドメインから200データずつランダムサンプリングする.



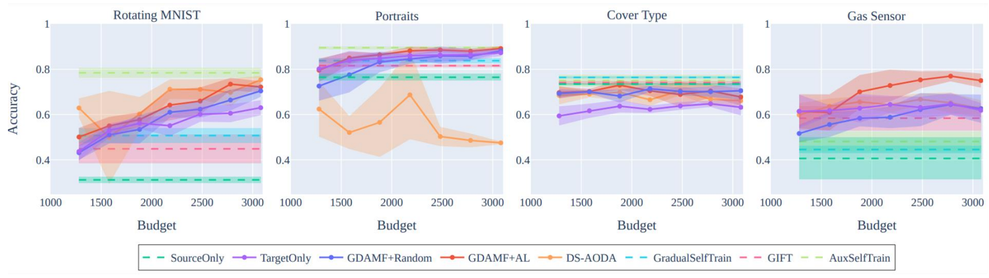
利用可能な中間ドメインが限られている場合でもGDAMFは有効





ドメイン毎に能動学習を実行することで, ランダムサンプリングよりも少ない予算で所定の正解率を達成できている

Multifidelityが機能しており,
コストとfidelityを考慮したクエリ数の設定がされている



GDAMFはどのデータセットにおいても安定して高い性能が得られている.

【まとめ、課題、今後の予定】

- 段階的ドメイン適応において, 既存手法では利用可能な中間ドメインが限られた場合は対応不可.
- ドメイン毎のクエリコストも考慮する手法として, Multifidelity能動学習を段階的ドメイン適応に応用するGDAMFを提案.
- 提案手法の有効性を4つの実データセットで確認した.
- 中間ドメインの数が多いほど, 巨大なモデルになることが課題
- 今後は, モデルの軽量化に加え, 中間ドメインを利用することの有効性について理論解析も行う予定.
- 本発表に関する論文は　https://arxiv.org/abs/2202.04359　から取得可能です。