FY2022/2022年度

High-Dimensional Statistical Modeling Team

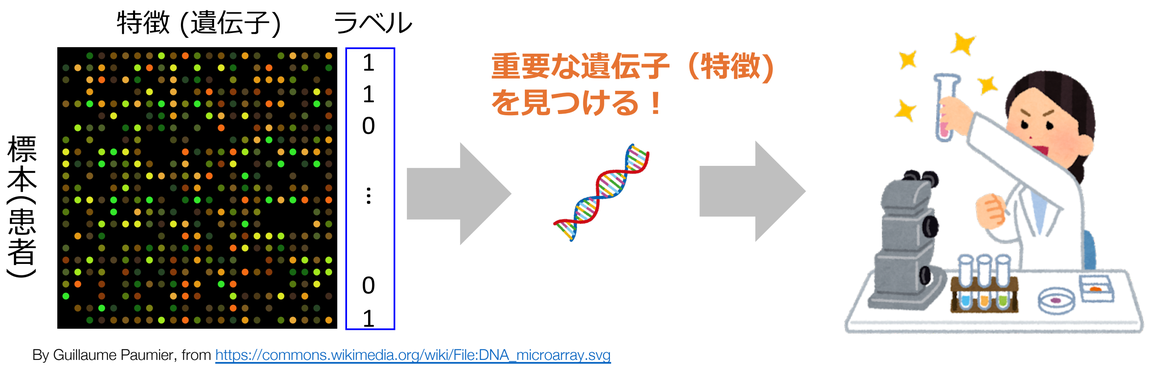
Makoto Yamada
高次元統計モデリングチーム　山田 誠

RIKEN Center for Advanced Intelligence Project

高次元統計モデリングチームの研究概要

目標

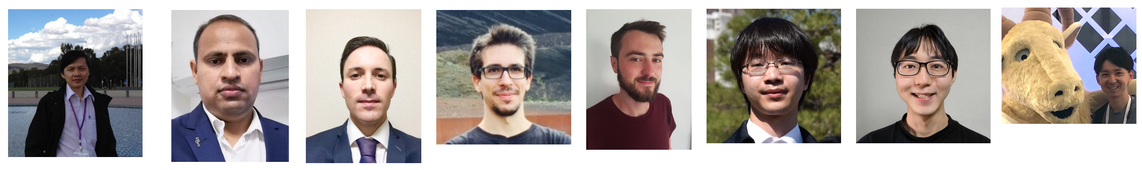
- 医療, 材料分野において**科学的発見をサポートする機械学習基盤**の構築
  - 重要な特徴をデータから精度よく容易に見つけられる手法の研究開発
  - 機械学習研究者以外でも容易に利用可能なソフトウェア開発
- 新規の科学的発見を容易にする基盤を確立し, **ヘルスケア, 材料, 農業等の分野で革新**を目指す
  - 医療費の削減 (個別化医療, 疾病予測)
  - 材料発見の大幅な効率化



By Guillaume Paumier, from https://commons.wikimedia.org/wiki/File:DNA_microarray.svg

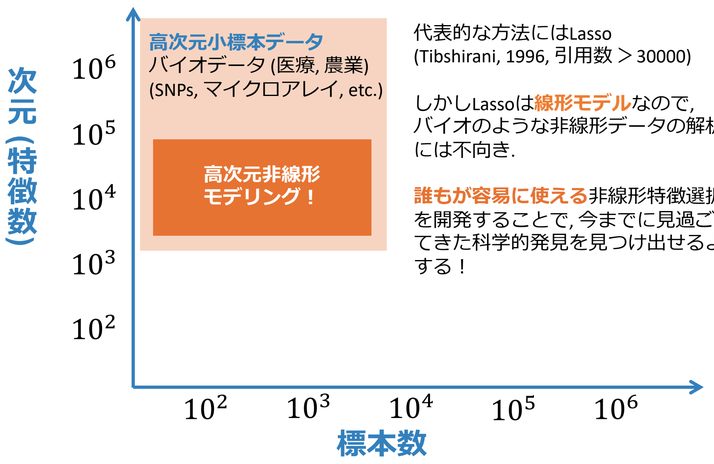
2022年度主要成果

- **特徴選択**
  - Knockoff filterに基づいた選択的推論 (AISTATS 2022)
  - Graph Neural Networkのための解釈方法GraphLimeの提案 (IEEE TKDE)
- **最適輸送（高次元データの類似度計算）**
  - 木構造Barycenter推定 (AISTATS 2022)
  - 1-Wasserstein距離の近似方法の提案 (TMLR 2022)
  - 高次元最適輸送手法の提案 (ECML 2022)
  - Word Mover's Distanceの再評価 (ICML 2022)
  - ロバストな最適輸送方法の提案 (ICLR 2023)
- **高次元・構造データの学習**
  - ヘシアン行列近似手法のimplicit differentiation への適応 (AISTATS 2023)
- **機械学習応用**
  - 細胞核画像分類のための手法の



主要成果1：高次元非線形特徴選択

高次元小標本データ



代表的な方法にはLasso
(Tibshirani, 1996, 引用数 > 30000)

しかしLassoは**線形モデル**なので,
バイオのような非線形データの解析
には不向き.

誰もが容易に使える非線形特徴選択手法を開発することで, 今までに見過ごされてきた科学的発見を見つけ出せるようにする！

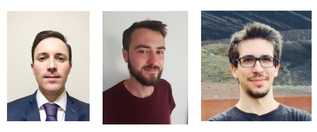
解釈性

- 少数特徴で高い予測性能
- 特徴の信頼度がわかる　（仮説検定）

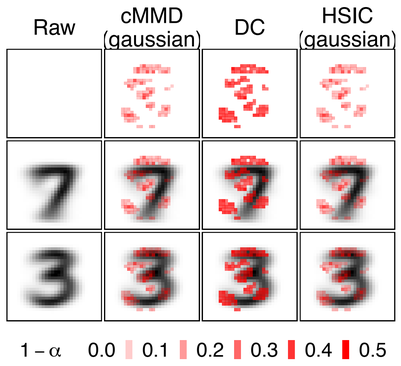
チャレンジ

- 非線形モデルは複雑　　（誰も利用しない）
- 非凸最適化が利用される（最適化が難しい）

カーネル選択的推論 (AISTATS 2022)



- Knockoff filterを初めてカーネル法(HSIC, MMD)に適用.
  - K個の特徴を選択
  - Knockoff filterを用いて検定を行う
- FDRのコントロールが可能
- Github
  PeterJackNaylor/knockoff-MMD-HSIC

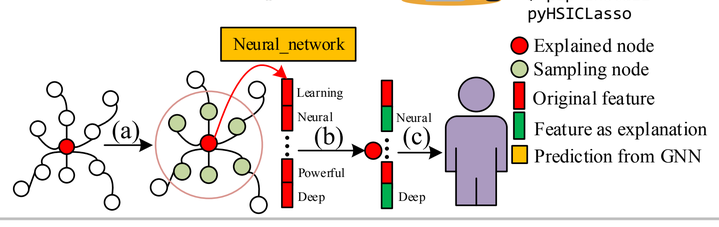


GraphLIMEの提案 (TKDE 2022)

- Graph Neural Network (GNN)の解釈手法であるGraphLIMEを提案
- アイディア: HSIC Lasso法をGNN解釈に適応

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d_+} \frac{1}{2}\left\|\bar{\boldsymbol{L}} - \sum_{k=1}^{d}\alpha_k \bar{\boldsymbol{K}}^{(k)}\right\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

pyHSICLasso
github.com/riken-aip/pyHSICLasso
$ pip install pyHSICLasso



主要成果2: 最適輸送

最適輸送の研究開発

我々のチームでは最適輸送の研究に取り組んでおり特に**最適輸送**の研究に取り組んでいる。

- 複数木に基づいた木構造Wasserstein距離 NeurIPS 2019
- 不均衡木構造最適輸送の研究 NeurIPS 2020, AISTATS 2021
- Neural Architecture Search (NAS)への応用 NeurIPS2021
- 異なるドメイン間の木構造最適輸送 AISTATS 2021
- 密度比推定に基づいた最適輸送 ECML 2021
- 木構造データにおけるBarycenter推定 AISTATS 2022
- 高次元データのための最適輸送 ECML 2022
- Word Mover's Distance (WMD)の再評価 ICML 2022
- 木構造最適輸送を用いたWasserstein距離の近似 TMLR 2022

木構造重心推定(AISTATS 2022)

- Tree Wasserstein distanceを利用したBarycenterの推定方法を提案
  - 提案法は凸最適化かつProjected gradient descentを使って高速に計算可能！

$$\mathbf{B} = \mathbf{w}_v \circ \begin{pmatrix} (\mathbf{I} - \mathbf{D}_1)^{-1}\mathbf{D}_2 \\ \mathbf{I} \end{pmatrix}, \quad (6)$$

$$f(\mathbf{a}) = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{B}\mathbf{a} - \mathbf{B}\mathbf{a}_i\|_1, \quad (7)$$

where $[\mathbf{a}_i]_k = \mu_i(v_{|\boldsymbol{V}_m|+k})$ and $[\mathbf{a}]_k = \mu(v_{|\boldsymbol{V}_m|+k})$. We define $\boldsymbol{A} = \{\mathbf{a} \in \mathbb{R}_+^{|\boldsymbol{V}_{\text{leaf}}|} \mid \|\mathbf{a}\|_1 = 1\}$. The FS-TWB problem can then be formulated as follows:

$$\bar{\mathbf{a}} \in \operatorname*{argmin}_{\mathbf{a}\in\boldsymbol{A}} f(\mathbf{a}). \quad (8)$$

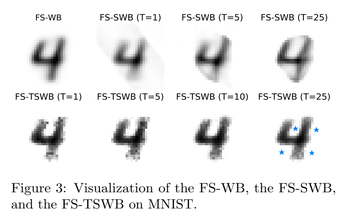


Figure 3: Visualization of the FS-WB, the FS-SWB, and the FS-TSWB on MNIST.

木構造Wasserstein距離を用いた1-Wasserstein距離の近似 (TMLR 2022)

任意のメトリックの**1-Wasserstein距離をL1距離で近似する方法(TWD)**を提案。

TWD: $W_{\mathcal{T}}(\mu,\nu) = \|\mathrm{diag}(\boldsymbol{w})\boldsymbol{B}(\boldsymbol{a}-\boldsymbol{b})\|_1$,

$\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{N_{\text{leaf}}}]$. 木のパラメータ

Lasso法に基づいた重み学習法を提案.

$$\widehat{\boldsymbol{w}} := \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d_+} \sum_{(i,j)\in\Omega} (d(\boldsymbol{x}_i, \boldsymbol{y}_j) - \boldsymbol{w}^\top \boldsymbol{z}_{i,j})^2 + \lambda\|\boldsymbol{w}\|_1$$

$$\boldsymbol{z}_{i,j} = \boldsymbol{b}_i + \boldsymbol{b}_j - 2\boldsymbol{b}_i \circ \boldsymbol{b}_j$$

$d(\boldsymbol{x}_i, \boldsymbol{y}_i)$ は1-Wasserstein距離の任意のメトリック.

1-Wasserstein距離 (WMD)よりも**数百倍高速かつ同等の性能**を得る方法を提案！

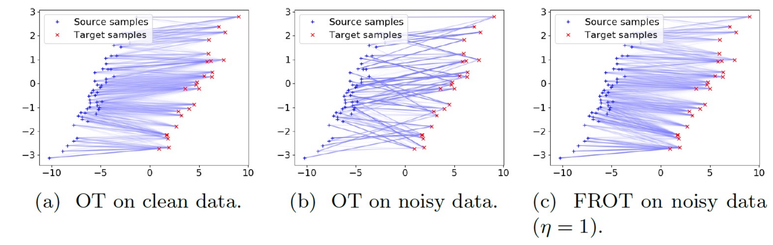
Methods	Twitter	BBCSport	Amazon
WMD (Sinkhorn)	0.675 ± 0.033	0.973 ± 0.015	0.903 ± 0.006
QuadTree	0.701 ± 0.027	0.970 ± 0.016	0.865 ± 0.001
qTWD	0.691 ± 0.028	0.967 ± 0.014	0.847 ± 0.036
Sliced-QuadTree	0.694 ± 0.017	0.970 ± 0.020	0.877 ± 0.001
Sliced-qTWD	0.697 ± 0.027	0.967 ± 0.014	0.887 ± 0.011
ClusterTree	0.683 ± 0.019	0.901 ± 0.056	0.873 ± 0.010
cTWD	0.699 ± 0.032	0.962 ± 0.016	0.878 ± 0.006
Sliced-ClusterTree	0.694 ± 0.010	0.929 ± 0.037	0.900 ± 0.011
Sliced-cTWD	0.700 ± 0.021	0.970 ± 0.018	0.905 ± 0.010

主要成果2：最適輸送 (続き)

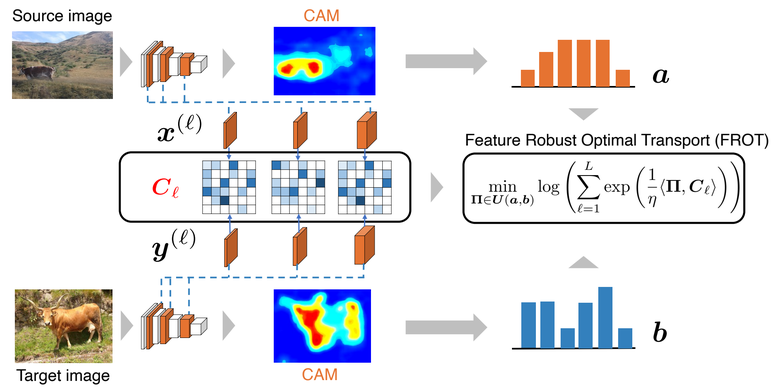
Feature Robust Optimal Transportの提案 (ECML 2022)

- Feature Robust Optimal Transport (FROT)

$$\mathrm{FROT}(\mu,\nu) = \min_{\boldsymbol{\Pi}\in\boldsymbol{U}(\mu,\nu)} \max_{\boldsymbol{\alpha}\in\boldsymbol{\Sigma}^L} \sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\sum_{\ell=1}^{L}\alpha_\ell c(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{y}_j^{(\ell)})$$



(a) OT on clean data. (b) OT on noisy data. (c) FROT on noisy data ($\eta = 1$).



Proposed method　　Baseline (CVPR 2020)

本年度まとめ

- 高次元非線形特徴選択手法の提案
- 最適輸送における重心(Barycenter)の超高速計算技術を開発
- 高次元データからの最適輸送手法を提案

お知らせ

高次元統計モデリングチームは2023/3月末で閉鎖します。
これまで6年間サポートいただきありがとうございました。
今後は研究の場を沖縄科学技術大学院大学に移し引き続き機械学習の研究活動を続ける予定です。上記研究テーマにご興味あれば是非 mlds@oist.jp までご連絡ください！