高次元統計モデリングユニット　山田　誠
High-dimensional Statistical Modeling Unit
Makoto Yamada

科学的発見のための機械学習技術の研究開発

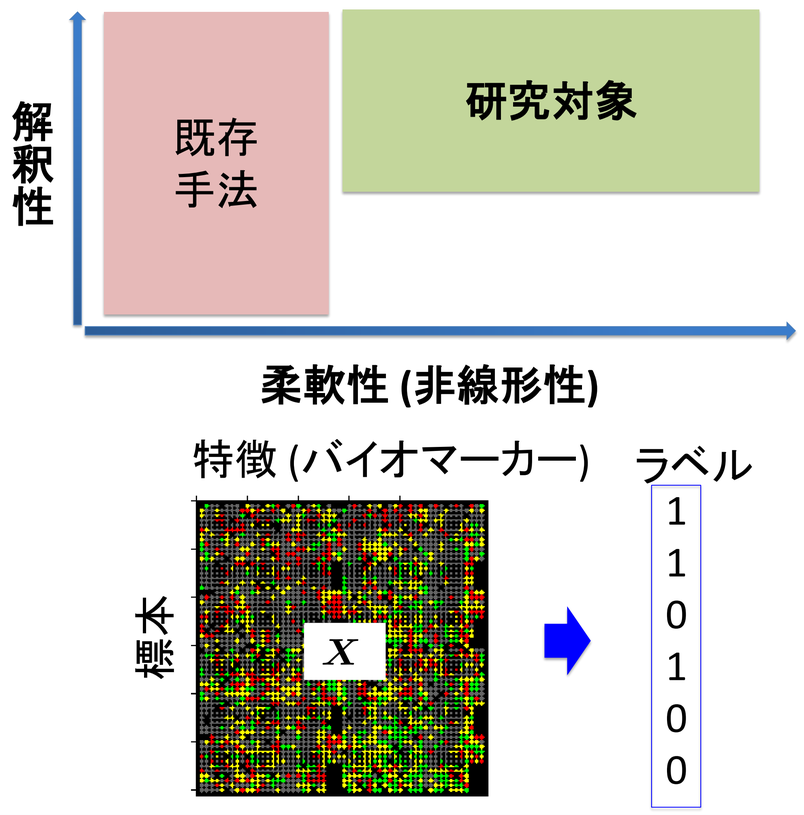
問題意識: バイオデータのような非線形データからは既存法では科学的発見ができない場合が多く, 新規の発見が見過ごされている可能性がある.

目標: 非線形データから自動的に新規の科学的発見をする枠組みの構築. 科学的発見を加速させる.

成果1: 100万次元を超える超高次元非線形データから, 中・小規模サーバを用いて自動で重要な特徴を選択する方法を開発.

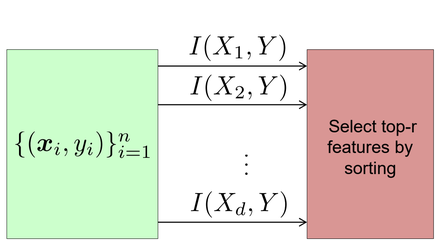
成果2: 高次元非線形データから, 個人の特徴を選択する手法(個別化特徴選択手法)の枠組みを提案.

成果3: 非線形関係のあるデータから, 自動で重要なデータセットを選択し, そのp値を計算する方法を世界で初めて開発.



Block HSIC Lasso [biorXiv 2019]

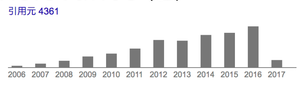
Sure Independence Screening (SIS法)



$$\max_{\beta\in\{0,1\}^d} \sum_{k=1}^d \beta_k I(X_k, Y), \quad \text{s.t.} \sum_{k=1}^d \beta_k = r$$

重複した特徴を選択してしまう☹

Minimum Redundancy Maximum Relevance 法 (mRMR)



$$\max_{\beta\in\{0,1\}^d} \sum_{k=1}^d \beta_k I(X_k, Y) - \sum_{k=1}^d \sum_{k=1}^d \beta_k \beta_{k'} I(X_k, X_{k'}), \quad \text{s.t.} \sum_{k=1}^d \beta_k = r$$

Block HSIC Lasso, mRMR法の凸最適化版

$$\max_{\alpha\geq 0} \sum_{k=1}^d \alpha_k \mathrm{HSIC}_b(f_k, y) - \frac{1}{2} \sum_{k,k'=1}^d \alpha_k \alpha_{k'} \mathrm{HSIC}_b(f_k, f_{k'}) - \lambda \|\alpha\|_1,$$

$$\mathrm{HSIC}_b(f_k, y) = \frac{B}{n} \sum_{\ell=1}^{n/B} \mathrm{HSIC}_v(f_k^{(\ell)}, y^{(\ell)}).$$

Github: https://github.com/riken-aip/pyHSICLasso

実験結果:

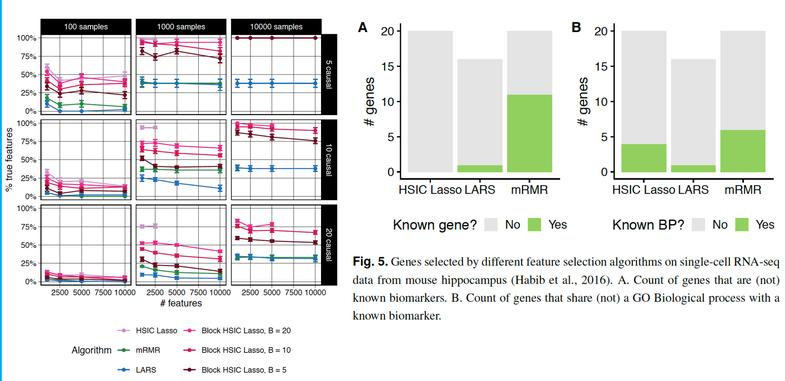


Fig. 5. Genes selected by different feature selection algorithms on single-cell RNA-seq data from mouse hippocampus (Habib et al., 2016). A. Count of genes that are (not) known biomarkers. B. Count of genes that share (not) a GO Biological process with a known biomarker.

理論: 大域的最適解がもとまる.

アルゴリズム: 数百万次元×数万サンプルからの特徴選択が可能. 少数特徴で高い予測精度.

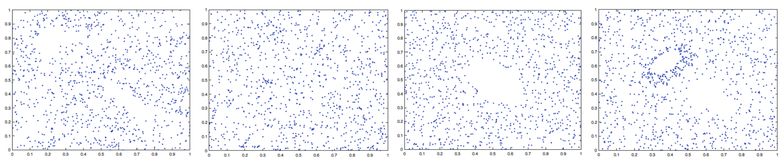
発表実績: BioRxiv 2019. ISMB 2019投稿.

応用展開: 骨髄性急性白血病データの解析(ヒト疾患モデル研究チーム), アトピー性皮膚炎データの解析(健康医療データAI予測推論開発ユニット).

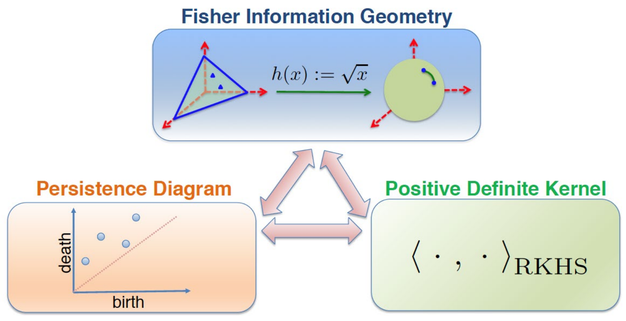
理論展開: 1. 理論解析. 2. グラフデータに対する特徴選択の問題に拡張.

Persistent Fisher Kernel [NeurIPS 2018]

位相データ解析 (TDA)



提案する枠組み



Persistence Fisher Kernel

$$k_{\mathrm{PF}}(\mathrm{Dg}_i, \mathrm{Dg}_j) := \exp\left(-t d_{\mathtt{FIM}}(\mathrm{Dg}_i, \mathrm{Dg}_j)\right)$$

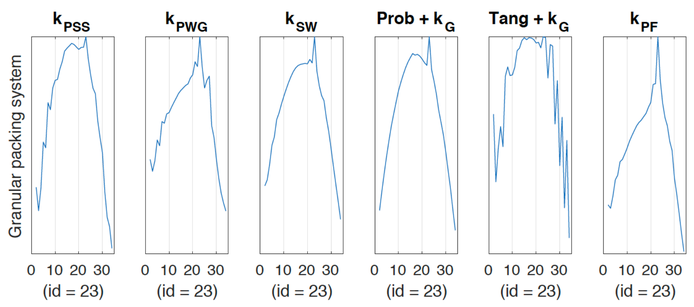
Algorithm 1 Compute $d_{\mathtt{FIM}}$ for persistence diagrams

Input: Persistence diagrams Dg_i, Dg_j, and a bandwith $\sigma > 0$ for smoothing
Output: $d_{\mathtt{FIM}}$

1: Let $\Theta \leftarrow \mathrm{Dg}_i \cup \mathrm{Dg}_{j\Delta} \cup \mathrm{Dg}_j \cup \mathrm{Dg}_{i\Delta}$ (a set for smoothed and normalized measures)
2: Compute $\bar{\rho}_i = \rho_{(\mathrm{Dg}_i \cup \mathrm{Dg}_{j\Delta})} \leftarrow \left[\frac{1}{Z} \sum_{u\in \mathrm{Dg}_i \cup \mathrm{Dg}_{j\Delta}} \mathtt{N}(x; u, \sigma I)\right]_{x\in\Theta}$
where $Z \leftarrow \sum_{x\in\Theta} \sum_{u\in \mathrm{Dg}_i \cup \mathrm{Dg}_{j\Delta}} \mathtt{N}(x; u, \sigma I)$
3: Compute $\bar{\rho}_j = \rho_{(\mathrm{Dg}_j \cup \mathrm{Dg}_{i\Delta})}$ similarly as $\bar{\rho}_i$.
4: Compute $d_{\mathtt{FIM}} \leftarrow \arccos\left(\langle \sqrt{\bar{\rho}_i}, \sqrt{\bar{\rho}_j} \rangle\right)$ where $\langle \cdot, \cdot \rangle$ is a dot product and $\sqrt{\cdot}$ is element-wise.

実験結果: 分類, 変化点検出

	MPEG7	Orbit
k_{PSS}	73.33 ± 4.17	72.38 ± 2.41
k_{PWG}	74.83 ± 4.36	76.63 ± 0.66
k_{SW}	76.83 ± 3.75	83.60 ± 0.87
Prob+k_G	55.83 ± 5.45	72.89 ± 0.62
Tang+k_G	66.17 ± 4.01	77.32 ± 0.72
k_{PF}	$\mathbf{80.00 \pm 4.08}$	$\mathbf{85.87 \pm 0.77}$



理論:パーシステント図間の類似度を精度よく測ることが可能.

発表実績: 難関国際会議NeurIPS 2018発表.

応用展開: 材料科学分野への応用.

理論展開: 1. 手法の高速化. 2. ベイズ最適化.

mmdInf [ICLR 2019]

Post Selection Inference: 選択したデータセットが統計的有意かどうかを検定する.

- H_0: $\sum_{s=1}^d \eta_s \widehat{D}(\boldsymbol{X}^{(s)}, \boldsymbol{Y}^{(s)}) = 0 \mid \mathcal{S}$ was selected,
- H_1: $\sum_{s=1}^d \eta_s \widehat{D}(\boldsymbol{X}^{(s)}, \boldsymbol{Y}^{(s)}) \neq 0 \mid \mathcal{S}$ was selected,

特徴

1. 二標本検定を複数データに拡張.
2. ノンパラメトリック.
3. 不完全U統計量Maximum Mean Discrepancy (MMD).

Theorem 1 (Lee et al., 2016) *Consider a stochastic data generating process* $\mathcal{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *If a feature selection event is characterized by* $\boldsymbol{A}\mathcal{Z} \leq \boldsymbol{b}$ *for a matrix* $\boldsymbol{A}$ *and a vector* $\boldsymbol{b}$ *that do not depend on* $\mathcal{Z}$, *then, for any fixed vector* $\boldsymbol{\eta} \in \mathbb{R}^d$,

データ選択 ↓ 切断正規分布の累積密度関数

$$F^{[V^-(\boldsymbol{A},\boldsymbol{b}), V^+(\boldsymbol{A},\boldsymbol{b})]}_{\boldsymbol{\eta}^\top \boldsymbol{\mu}, \boldsymbol{\eta}^\top \boldsymbol{\Sigma} \boldsymbol{\eta}} (\boldsymbol{\eta}^\top \mathcal{Z}) \mid \boldsymbol{A}\mathcal{Z} \leq \boldsymbol{b} \sim \mathrm{Uni}(0, 1),$$

where $F^{[v,w]}_{t,u}(\cdot)$ *is the cumulative distribution function of the uni-variate truncated normal distribution with the mean* t, *the variance* u, *and the lower and the upper truncation points* v *and* w, *respectively. Furthermore, using* $\boldsymbol{c} := \frac{\boldsymbol{\Sigma}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top \boldsymbol{\Sigma} \boldsymbol{\eta}}$, *the lower and the upper truncation points are given as*

$$V^-(\boldsymbol{A}, \boldsymbol{b}) := \max_{j:(\boldsymbol{A}\boldsymbol{c})_j < 0} \left\{ \frac{b_j - (\boldsymbol{A}\boldsymbol{z})_j}{(\boldsymbol{A}\boldsymbol{c})_j} \right\} + \boldsymbol{\eta}^\top \boldsymbol{z}, \quad (1)$$

$$V^+(\boldsymbol{A}, \boldsymbol{b}) := \min_{j:(\boldsymbol{A}\boldsymbol{c})_j > 0} \left\{ \frac{b_j - (\boldsymbol{A}\boldsymbol{z})_j}{(\boldsymbol{A}\boldsymbol{c})_j} \right\} + \boldsymbol{\eta}^\top \boldsymbol{z}. \quad (2)$$

実験結果:

Table 1: Post selection inference experimental results for real-world datasets. The average TPR and FPR over 200 trials are reported.

Datasets	d	n	Linear-Time		Block $B=10$		Block $B=20$		Block $B=50$		Incomplete $r=0.5$		Incomplete $r=5$		Incomplete $r=10$	
			TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Diabetis	8	768	0.05	0.06	0.20	0.04	0.40	0.12	0.46	0.18	0.13	0.06	0.52	0.07	0.65	0.10
Wine (White)	11	4898	0.13	0.06	0.26	0.01	0.56	0.04	0.68	0.09	0.32	0.06	0.71	0.06	0.79	0.06
Wine (Red)	11	1599	0.08	0.06	0.26	0.04	0.37	0.09	0.49	0.15	0.15	0.06	0.51	0.07	0.61	0.07
Australia	7	690	0.08	0.06	0.36	0.06	0.47	0.11	0.65	0.21	0.16	0.05	0.66	0.07	0.79	0.09

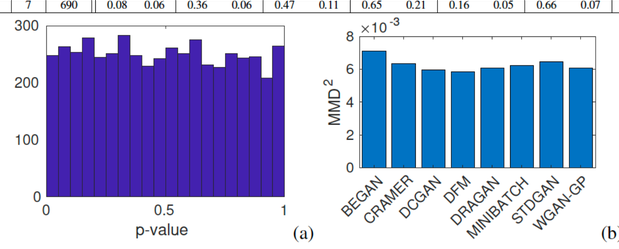


Figure 4: (a) Histogram of p-values over 5000 runs. (b) Averaged incomplete MMD scores.

理論: 標本数が十分に大きければ, Type I errorを理論的にコントロールできる.

アルゴリズム: シンプルであり, 標本数が大きい場合にも高速に計算可能.

発表実績: 難関国際会議ICLR 2019採択.

応用展開: 薬剤誘発による細胞の形態変化解析に利用するため, 実証実験を開始. Incomplete MMDを用いたアルゴリズムの開発.

理論展開: 高次元小標本の場合への拡張.