研　究　報　告　書

「科学的発見のための非線形機械学習技術の創生」

研究タイプ：通常型
研究期間： 2016 年 12 月～2020 年 3 月
研 究 者： 山田　誠

1．　研究のねらい

マテリアルズインフォマティクス，バイオインフォマティクス，ヘルスケア等の分野において，計測機器の高度化や実験手法の発展により**大量かつ複雑なデータ**を得られるようになっている．しかし，上記分野においては，新規の科学的発見をすることが非常に重要なタスクであるが，近年では新薬開発に数百億の費用がかかる等，データ増加だけでなく新規の重要な物質や化合物を見つけるための期間や費用も増大している．そのため，科学的発見を効率化し，医薬品や製品をいち早く消費者の元へ届けるだけではなく，製品価格の上昇を抑えることが学術界のみならず産業界においても重要課題となっている.このような背景のもと，人工知能や機械学習といった統計的な手法がデータ処理に積極的に用いられている．特に，**機械学習**は近年飛躍的な発展を遂げており，学術界・産業界から大きな注目を浴びている．本研究の狙いは，材料科学や創薬科学等の新規の科学的発見が非常に重要な分野において，大量の候補から効率よく目的となる材料や化合物を自動的に見つけ出す機械学習技術基盤を構築することである．

私が提案する研究プロジェクトでは，材料の候補物質選択や個別創薬に関連する高次元データ処理課題を抽出し，その課題を機械学習の問題として定式化する．具体的には，グラフ構造や非線形構造を持つ高次元大規模データベースからの特徴選択のデータ処理課題を軸に研究を展開していく．本プロジェクトでは，スパースモデリングを用いて上記の課題を定式化し，新規の科学的発見手法の基礎理論を確立する．さらに，高精度かつスケーラブルな機械学習ソフトウェア基盤を構築し，材料科学や創薬科学においてパラダイムを生み出すことを目指す．

2．　研究成果

（1）概要

本さきがけにおいては，高次元小標本に対する，非線形高次元統計モデリングに関する基礎研究およびその応用研究に取り組んだ．その結果，非線形特徴選択の方法論において，数十万次元といった超高次元データから入出力間に非線形関係のある特徴を選択する方法を開発できた．さらに，小規模サーバーで超高次元データから特徴選択できるように HSIC Lasso を拡張した Block HSIC Lasso を提案した．本拡張により，従来は扱うことのできなかった GWAS データから効率的に非線形関係のある特徴を選択できるようになった．そして,白血病，アトピー性皮膚炎の発現量データにこれらの開発した特徴選択手法を用いることで,白血病に関しては有望な治療標的をいくつか見いだすことに貢献した．

超高次元特徴選択の貢献に加え，本さきがけを通して非線形選択的推論の方法を世界に先駆けて提案することができた．具体的には，重要な非線形手法の代表的な方法であるカーネル法に基づいた非線形選択的推論の枠組みを確立した．選択的推論とは，特徴選択後で選択した特徴が統計的に優位かどうかを検定する方法論のことである．本貢献により，既存の独

立性検定，二標本検定，適合度検定において選択的推論が可能となった．今後，提案する枠組みが拡張されることで，新規の科学的発見を加速できると確信している．

（2）詳細

研究テーマ A「非線形特徴選択手法のアルゴリズム開発」

（A-1）高次元非線形特徴選択に関する研究

大規模分散処理を利用して数十万次元といった超高次元データから入出力間に非線形関係のある特徴を選択する方法を開発した．さらに，小規模サーバーで数十万次元の大きさの問題を解けるように HSIC Lasso を拡張した Block HSIC Lasso を提案した．本拡張により，従来は扱うことのできなかった GWAS データから効率的に非線形関係のある特徴を選択できるようになった（表1）．

表1．大規模データにおける実験結果．

データ	特徴数	サンプル数	Raw	LARS	HSIC Lasso
RA	352,773	3,451	0.671 ± 0.002	0.572 ± 0.002	**0.767 ± 0.004**
T1D	352,853	3,443	0.671 ± 0.006	0.569 ± 0.004	**0.788 ± 0.002**
T2D	353,046	3,456	0.609 ± 0.004	0.565 ± 0.005	**0.675 ± 0.003**

さらに，提案した Block HSIC Lasso の Python コード(pyHSICLasso)を Github で配布する等，機械学習以外の研究者も容易に使えるようにした．これらの研究成果を，トップジャーナルである IEEE TKDE および Bioinformatics 氏に報告したところ，TKDE に関しては出版後 popular Documents の上位に 1 年以上連続でランクされる等非常に高い注目を集めた． バイオデータのような数十万次元といった高次元データから非線形性を利用して特徴選択できるソフトウェアは非常に少ないため，今後本ソフトウェアが普及することで，従来見つからなかった発見が可能となることを期待している．HSIC Lasso のアルゴリズム研究開発に加え，HSIC Lasso の理論的性質（Consistency, Support recovery)を明らかにした．本研究成果は，AISTATS 2020 に報告予定である．

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

pyHSICLasso

github.com/riken-aip/pyHSICLasso

```
$ pip install pyHSICLasso
```

図1．pyHSICLasso パッケージ．

（A-2）非線形選択的推論の確立

バイオデータのような入出力間に非線形関係のあるデータから，重要な非線形特徴を選択できるカーネル法に基づいた非線形選択的推論の枠組みを確立した．選択的推論とは，特徴選

択後で選択した特徴が統計的に優位かどうかを検定する方法論のことである．これにより，既存の独立性検定，二標本検定，適合度検定において選択的推論が可能となった．以下に，本さきがけで提案した方法をまとめる．

- **選択的独立性検定**：Hilbert-Schmidt Independence Criterion(HSIC)を用いた選択的推論手法(hsicInf)を開発．（AISTATS 2018）
- **選択的二標本検定**：Maximum Mean Discrepancy (MMD)を用いた選択的推論手法(mmdInf)を開発．(ICLR 2019)
- **選択的適合度検定**：Kernel Stein Discrepancy (KSD)を用いた選択的推論手法(ksdInf)を開発．(NeurIPS 2019)

図2に人工データにおける True Positive Rate (TPR)および False Positive Rate (FPR)の結果を示す．この結果から，線形手法である larInf は線形データには有用であることがわかるが(図2-(a))，一方で非線形データにおいては全く重要な特徴を検出できていないことがわかる(図2-(b))．提案法である hsicInf はそのような非線形データにおいて重要な特徴を選択できている．

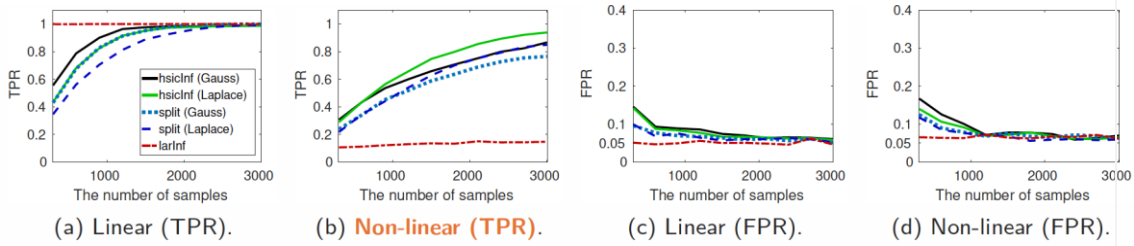

図2　hsicInf の結果．

研究テーマ B「機械学習アルゴリズムの開発」

(B-1) 構造化データからの非線形特徴選択法の開発

HSIC Lasso を拡張しグラフデータからの非線形特徴選択法を開発した．具体的には，HSIC Lasso が弱教師有り学習の設定にも容易に拡張できる特性を利用して，各ノードに高次元のデータがのっているグラフデータからの特徴選択手法を提案した．本研究成果は AAAI 2020 に採録され，単純な方法ながら高い評価を受けた．

(B-2) トポロジカルデータ解析のためのカーネル法の研究開発

近年，トポロジー情報を機械学習に利用する研究が盛んに行われており，トポロジカルデータ解析(TDA)と呼ばれている．TDA においては，パーシステント図を用いたデータ解析が主流であり，そのためパーシステント図間の類似度を測ることが非常に重要である．そこで，我々は，パーシステント図が非負データであることに注目し，Fisher 情報量に基づいた新規のカーネルを提案した．さらに，本手法を変化点検出等のタスクに適用したところ，非常に高い精度を得ることができた．さらには，提案したカーネルを用い,トポロジー情報を用いたベイズ最適化の枠組みを提案した．

研究テーマ C「実応用」
機械学習の応用研究としては，以下の4つのプロジェクトと共同で研究を実施した.

(C-1) HSIC Lasso を用いたうつ病予測の研究
HSIC Lasso を利用したことで，従来法よりも高い精度でうつ病の兆候を予測することができるようになった．本結果は，Translational Psychiatry に採録された.

(C-2) 急性白血病のための治療法の探索
白血病のデータから HSIC Lasso を利用することで治療標的を見つけ，その見つけ出した遺伝子を阻害する薬を使ってケミカルスクリーニングを実施した．そうしたところ，白血病を殺すことのできる治療標的を複数発見することができ，その知見を元にヒト化マウスを用いて実験をしているところである．本成果は，ジャーナル投稿に向けて準備しているところである.

(C-3) アトピー治療法の探索
アトピーデータ（発現量）から HSIC Lasso を用いて複数の特徴を選択し，選択した特徴から創薬に利用できそうな遺伝子を絞り込んでいるところである．今後は，選択した特徴からペプチドを作り，マウスを用いて評価する予定である.

(C-4) グラフェン画像処理
グラフェン画像から自動でグラフェンの枚数を推定する問題に関して，物体検出アルゴリズムの U-net を利用することで，少しの訓練を受けた人間と同等以上の検出結果を出すことができた．本成果は NPJ Computational Materials (Impact factor 9.2)に採録された.

3． 今後の展開
さきがけ研究において，高次元データから非線形特徴選択のできる方法を複数提案し，特に HSIC Lasso がバイオや医療データにおいて有用であることを実験から検証した．その一方で，機械学習手法単体における新規の科学的発見が非常に難しいということも今回の研究を通じてわかった．そのため，データからのみの発見を目指すだけではなく，評価系も含め治療標的探索プロセス全体を最適化する人間参加型機械学習基礎技術の研究が重要だと考えている.

4． 自己評価
研究目標は概ね達成されたと考えている．特に，非線形特徴選択の方法論においては，従来予定していたよりもはるかに大きな成果を得られることができた．また，機械学習系の難関国際会議に10本以上の成果を出せる等，十分に大きな成果を出せたと考えている．一方で，社会実装に関連する新規科学的発見に関しては，期間中に何かしらの成果を公表できると考えていたが，自分が当初予定しているよりも時間がかかってしまっている．これは機械学習の研究に比べサイエンスの研究の方が実験や共同研究に時間がかかることが主な理由であるが，こちらは工夫

の仕方によってはより効率化できるのではないかと考えている.

研究成果の科学技術及び社会・経済への波及効果に関しては，現在，白血病やアトピー性皮膚炎の新規治療薬の共同研究を進めているところであり，概ね順調に進んでいる．そのため，これらの成果が世にでることで達成されると考えている．また，今回開発した pyHSICLasso のパッケージに関しては，すでに公開しており，多くの研究者が利用を始めている．そのため，方法開発の成果は直接見えにくいが，着実にサイエンスの研究を加速することに貢献していることを実感している.

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. **Makoto Yamada**, Yuta Umezu, Kenji Fukumizu, Ichiro Takeuchi: Post Selection Inference with Kernels. AISTATS 2018: 152-160.
2. **Makoto Yamada**, Jiliang Tang, Jose Lugo-Martinez, Ermin Hodzic, Raunak Shrestha, Avishek Saha, Hua Ouyang, Dawei Yin, Hiroshi Mamitsuka, Süleyman Cenk Sahinalp, Predrag Radivojac, Filippo Menczer, Yi Chang: Ultra High-Dimensional Nonlinear Feature Selection for Big Biological Data. IEEE Trans. Knowl. Data Eng. 30(7): 1352-1365 (2018).
3. **Makoto Yamada**, Denny Wu, Yao-Hung Hubert Tsai, Hirofumi Ohta, Ruslan Salakhutdinov, Ichiro Takeuchi, Kenji Fukumizu: Post Selection Inference with Incomplete Maximum Mean Discrepancy Estimator. ICLR (Poster) 2019.
4. Héctor Climente-González, Chloé-Agathe Azencott, Samuel Kaski, **Makoto Yamada**: Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. Bioinformatics 35(14): i427-i435 (2019).
5. Jen Ning Lim, **Makoto Yamada**, Bernhard Schölkopf, Wittawat Jitkrittum: Kernel Stein Tests for Multiple Model Comparison. NeurIPS 2019: 2240-2250.
6. Benjamin Poignard, **Makoto Yamada**, Sparse Hilbert-Schmidt Independence Criterion Regression. AISTATS 2020, in press.
7. Jenning Lim, **Makoto Yamada**, Wittawat Jitkrittum, Yoshikazu Terada, Shigeyuki Matsui, Hidetoshi Shimodaira. More Powerful Selective Kernel Tests for Feature Selection AISTATS 2020, in press.

(2)特許出願

研究期間累積件数： 0 件(公開前の出願件名については件数のみ記載)

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

1. Ryoma Sato, **Makoto Yamada**, Hisashi Kashima: Approximation Ratios of Graph Neural Networks for Combinatorial Problems. NeurIPS 2019: 4083-4092
2. Tam Le, **Makoto Yamada**, Kenji Fukumizu, Marco Cuturi: Tree-Sliced Variants of Wasserstein Distances. NeurIPS 2019: 12283-12294
3. Tam Le, **Makoto Yamada**: Persistence Fisher Kernel: A Riemannian Manifold Kernel for

Persistence Diagrams. NeurIPS 2018: 10028–10039

4. **Makoto Yamada**, Wenzhao Lian, Amit Goyal, Jianhui Chen, Kishan Wimalawarne, Suleiman A. Khan, Samuel Kaski, Hiroshi Mamitsuka, Yi Chang: Convex Factorization Machine for Toxicogenomics Prediction. KDD 2017: 1215–1224

5. **Makoto Yamada**, Koh Takeuchi, Tomoharu Iwata, John Shawe-Taylor, Samuel Kaski: Localized Lasso for High-Dimensional Regression. AISTATS 2017: 325–333