様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１１　日現在

機関番号：　８２４０１
研究種目：　若手研究(B)
研究期間：　2016～2017
課題番号：　１６Ｋ１６１１４

研究課題名（和文）超高次元データに対する非線形解析手法の研究開発

研究課題名（英文）Developing Nonlinear Feature Selection Algorithm for Ultra High-Dimensional Data

研究代表者
山田　誠（Yamada, Makoto）

国立研究開発法人理化学研究所・革新知能統合研究センター・ユニットリーダー

研究者番号：００５８１３２３

交付決定額（研究期間全体）：（直接経費）　　1,500,000円

研究成果の概要（和文）：本研究を通して，100万次元を超える超高次元特徴から数時間で非線形性の特徴を選択できるアルゴリズムを世界で初めて開発した．　さらに，機械学習分野外の研究者が開発したプログラムを利用できるように，提案法(HSIC Lasso)をPythonで実装し（pyHSICLasso），Githubにてソースコードを公開することに加え，利用者が容易にプログラムをインストールできるようにした．本研究成果はデータマイニングの難関ジャーナルであるIEEE Transactions on Knowledge and Data Engineering (TKDE)に投稿し採録された．

研究成果の概要（英文）：We have developed a nonlinear feature selection algorithm for ultra-high dimensional data (more than 1 million features with tens of thousand data samples). To the best of our knowledge, this is the first algorithm that scales to such data. Moreover, for non-machine learning researchers, we developed a python package "pyHSICLasso" and distributed the code through Github. Now, we can install the software using "pip install pyHSICLasso". Finally, our research paper entitled "Ultra High-Dimensional Nonlinear Feature
Selection for Big Biological Data" was accepted to a top-tier data mining journal IEEE Transactions on Knowledge and Data Engineering (TKDE).

研究分野：　機械学習

キーワード：　特徴選択

１．研究開始当初の背景

バイオインフォマティクス，コンピュータビジョン，自然言語処理，音声信号処理，Web マイニング,センサーデータ処理等の応用分野においては，高次元データから解釈可能な特徴を選択する手法(特徴選択) や画像認識等の分類問題において重要な特徴を生成する手法(特徴抽出) が非常に重要であり，これまでに様々な手法が開発されている．特に，標本数が大量に利用可能である一般画像認識や音声認識においては，特徴抽出手法の一種であるディープラーニング技術を用いることで性能が近年格段に向上している．その一方で，バイオインフォマティクスやヘルスケアといった分野では，大量の学習標本を集めることが難しい一方で，特徴数(例: 遺伝子変異箇所数: SNP) が数百万～数千万となることや，特徴を解釈し新しい科学的発見をすることが重要であるため，特徴抽出手法であるディープラーニング技術を直接応用することが難しい．また遺伝子解析においては，データが非線形な振る舞いをすることが多いため，非線形モデルを用いてデータ解析するのが自然であるが，非線形モデルの推定は難しいため現状は線形モデルを用いるのが主流である．特に百万次元以上の超高次元データから非線形関係のある特徴を選択するような研究に関しては，国際的に見ても皆無である． このような背景のもと，申請者は情報理論に基づく高次元データ解析アルゴリズムの研究開発を2011年から4 年間に渡り，国内大学や研究機関のみならずカーネギーメロン大学やインディアナ大学の海外の大学の研究者と共同で行ってきた．

２．研究の目的

本研究プロジェクトでは，数百万～数千万次元×数万標本からなる超高次元データからの非線形特徴選択手法を開発することである．具体的には，機械学習手法(HSIC Lasso)を百万次元の超高次元データを扱えるように拡張し，その後，提案手法を用いてバイオインフォマティクスやヘルスケアといった応用分野において新規の科学的発見を目指す．また，提案手法のソフトウェアを分野外の研究者が使いやすいように整備し，マテリアルズインフォマティクス等の他の分野の研究者との共同研究を加速し，提案法の有用性を示す．

３．研究の方法

超大規非線形特徴選択アルゴリズムの開発

(1) 大規模データへの対応: HSIC Lasso を大規模超高次元データに適応できるように拡張する．具体的には, HSIC Lasso の類似行列を少ないメモリ量で精度良く近似できるアルゴリズムを提案する．

(2) 超高次元データへの対応: 数百万次元のような超高次元データでは，メモリ量に加えて計算量が爆発的に大きくなることが予想される．このような場合には，一度に複数の特徴を選択するような方法ではなく，特徴を一つずつ順番に選択していくアプローチ(Forward selection) が有効である．そこで, HSIC Lasso の問題を Forward selection の問題として定式化し，さらにHadoop やSpark のような大規模分散処理フレームワークを用いて高速に特徴を選択できる手法を開発する．

(3) SNP データからの科学的発見や予測精度の高いツールの構築: SNP データは数百万～数千万次元のベクトルで表現されることが多いため，先行研究では処理量の少ない線形手法が用いられてきた．したがって，非線形性を扱える提案手法を用いることで，線形手法では見つけられない関係を

発見できる可能性がある. 本研究では, まずは数十万次元×数千標本程度(10 ギガバイト程度) の公開データを用いて提案法の有効性を検討する. そして, 医学部あるいは医療機関と協力して実際のゲノム配列データに適用し, 医療分野において予測精度の高いツールの構築を目指す.

(4) グラフ分類: バイオインフォマティクスの分野では, 標本が分子グラフで表現されることがあるため, グラフ間の類似性を図るためにグラフカーネルを用いることが一般である. 例えば, グラフレットカーネルでは, グラフデータから次数n が1-4の部分グラフ(グラフレット) を全てカウントした数百万次元のベクトルを準備し, それらの内積で類似度を計算する. しかし, 一般的にグラフカーネルには計算量が大きいという問題がある. その他の方法としては, 頻出パタン(サブグラフ) マイニングがあるが, この方法はサブグラフマイニングしたあとに有意性を検討するため2ステップの処理が必要となる. そこで, 提案法を用いて超高次元ベクトルから分類に重要なグラフレットを数十個直接選択できれば, 全てのグラフレットを計算する必要が無くなり分類時のグラフレットの計算量を大幅に削減できる.

４．研究成果

非線形特徴選択アルゴリズムの大規模データへの対応に関しては, HSIC LassoにNystrom近似を用いることで,アルゴリズムに必要なメモリ量を1000分の1程度に圧縮することに成功した. さらに, 超高次元データへの対応に関してはForward Selectionと大規模分散処理フレームワーク (Apache Spark)を利用することで, 100万次元×1万標本のデータでも効率よく処理できるフレームワークを確立した. 具体的には, 100万次元×1万標本のデータから, 数時間で入力と出力間に非線形性のあるような特徴を選択できることが可能となった. このように, 100万次元を超える超高次元特徴から数時間で非線形性の特徴を選択できるアルゴリズムは世界初である.

実応用に関しては, 提案した超大規模特徴選択アルゴリズムを前立腺癌の予測タスク (27万次元400サンプル)および酵素の識別タスク(106万次元1万5千サンプル)に適用した. 従来法では高い予測精度を得るために数千特徴が必要であったが, 提案法では数十特徴のみで従来法と同等以上の精度が得られることを確認した. 従来は高い性能を得るために多くの特徴が必要であり, モデルの解釈が難しかったが, 提案手法は数十特徴のみで高い性能が得られるため, モデルの解釈がしやすく大変有用であることがわかった. そして, これらの成果をまとめた論文は, データマイニングの難関ジャーナルであるIEEE Transactions on Knowledge and Data Engineering (TKDE)に採録された.

さらに, 国内外で多くの講演を実施し, 手法の良さを広めることで共同研究者を募集した. その結果, 東北メディカルバンクとの共同研究に発展した. また, 幅広いユーザーが研究成果を利用できるようにPythonでプログラムを実装し, Githubにてソースコードを公開することに加え, 利用者が容易にプログラムをインストールできるようにした.

最後に, 今回の研究成果に加え, 大規模非線形特徴選択アルゴリズム開発で重要な技術であるスパースモデリングを利用した研究成果を, 難関国際会議であるIJCAI, ACL, KDD及びAISTATSにて報告した.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2 件）

① Makoto Yamada, Jiliang Tang, Jose Lugo-Martinez, Ermin Hodzic, Raunak Shrestha, Avishek Saha, Hua Ouyang, Dawei Yin, Hiroshi Mamitsuka, Cenk Sahinalp, Predrag Radivojac, Filippo Menczer, Yi Chang
Ultra High-Dimensional Nonlinear Feature Selection for Big Biological Data. IEEE Transactions on Knowledge and Data Engineering, vol. 30, Issue:7,1352-1365,July 2018.査読有.

② Yi Chang, Makoto Yamada, Antonio Ortega, Yan Liu
Lifecycle Modeling for Buzz Temporal Pattern Discovery, TKDD 11(2) 20:1-20:24 (2016). 査読有.

〔学会発表〕(計 6件)

① Makoto Yamada, Koh Takeuchi, Tomoharu Iwata, John Shawe-Taylor, Samuel Kaski
Localized Lasso for High-Dimensional Regression.AISTATS 2017: 325-333. 査読有.

② 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. 査読有.

③ Zornitsa Kozareva, Makoto Yamada
Which Tumblr Post Should I Read Next? ACL (2) 2016. 査読有.

④ Junning Gao, Makoto Yamada, Samuel Kaski, Hiroshi Mamitsuka, Shanfeng Zhu
A Robust Convex Formulation for Ensemble Clustering. IJCAI 2016 1476-1482. 査読有.

⑤ Yi Chang, Jiliang Tang, Dawei Yin, Makoto Yamada, Yan Liu:
Timeline Summarization from Social Media with Life Cycle Models. IJCAI 2016 3698-3704. 査読有.

⑥ Tomoharu Iwata, Makoto Yamada:
Multi-view Anomaly Detection via Robust Probabilistic Latent Variable Models.NIPS 2016: 1136-1144. 査読有.

〔その他〕
ホームページ等
https://github.com/riken-aip/pyHSICLasso

６．研究組織
研究代表者
山田 誠 (YAMADA Makoto)
国立研究開発法人理化学研究所・革新知能統合研究センター・ユニットリーダー

研究者番号：00581323