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

令和　３ 年　 ６ 月 ２２ 日現在

機関番号：１３９０３
研究種目：基盤研究(A)（一般）
研究期間：2017～2020
課題番号：１７Ｈ００７５８

研究課題名（和文）ビッグデータ駆動型科学のための仮説生成・検証法開発と材料，生物，医療分野での実証

研究課題名（英文）Methods for selecting and testing hypothesis in big data-driven science and its demonstration in materials, biology, and medicine

研究代表者
竹内　一郎（Takeuchi, Ichiro）

名古屋工業大学・工学（系）研究科（研究院）・教授

研究者番号：４０３３５１４６

交付決定額（研究期間全体）：（直接経費）　　28,000,000円

研究成果の概要（和文）：さまざまな科学研究の分野で研究対象に関する膨大なデータが計測できるようになった．このようなデータに基づいて科学的発見を目指すアプローチはデータ駆動型科学と呼ばれている．データ駆動型科学では，データに基づいて仮説を選択するが，データにとって都合のよい仮説が誤って選択されるリスクがあり，適切に信頼性評価を行わなくてはならない．本研究では，選択的推論と呼ばれる技術を用いて，材料，生物，医療分野におけるデータ駆動型仮説の信頼性評価を行う方法を確立し，その実証を行った．

研究成果の学術的意義や社会的意義
研究対象から得られるデータに基づいて科学的発見を目指すアプローチはデータ駆動型科学と呼ばれ，さまざまな分野で有望視されている．しかしながら，データから仮説を選択する際に選択バイアスが生じてしまい，特に，誤った意思決定が重大なリスクとなる分野においては，データ駆動型仮説の信頼性評価が不可欠である．本研究ではデータ駆動型仮説の信頼性評価を行うための方法論を確立し，これをさまざまな分野で実証した．本研究の成果は健在なデータ駆動型科学の発展に寄与するものである．

研究成果の概要（英文）：In various fields of scientific research, it has become possible to measure vast amounts of data about the research subject. The approach to scientific discovery based on such data is called data-driven science. In data-driven science, hypotheses are selected based on the data, but there is a risk that hypotheses that are over-fitted to the data may be selected incorrectly, and reliability evaluation of the data-driven hypotheses must be conducted appropriately. In this study, we established and demonstrated a method for evaluating the reliability of data-driven hypotheses in the fields of materials, biology, and medicine using a technique called selective inference.

研究分野：機械学習

キーワード：機械学習　統計科学　材料科学　生物化学　医療科学

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

様々な計測技術の飛躍的な発展により，多くの科学分野にて高次元の網羅的データ（ビッグデータ）を取得できるようになった．ビッグデータを用いて科学的発見を目指すアプローチは，理論，実験，計算に次ぐ第４の科学的アプローチとして有望視されている．ビッグデータを用いることで，知識や経験のみからは想起できないような複数の要因に基づく複雑な仮説（複数要因仮説）を生成できる可能性がある．例えば，生物や医療の分野では，複数の遺伝的要因の組み合わせにより様々な生命現象が起こると考えられているが，知識や直感のみからそのような複雑な仮説をたてるのは難しい．また，材料分野では，複数原子の相対的配置によって材料の物性が定まるが，複数原子の影響を物理的な知見のみに基づいて予測するのは困難である．一方，高次元データを網羅的に探索して選び出された仮説が，データのばらつきによる偶然の産物であるのか，信頼の高い統計的に有意なものであるのかを適切に判断するのは難しい．仮説がデータに基づいて生成された場合，仮説検定の結果に「選択バイアス」と呼ばれる偏りが生じ，従来の仮説検定による評価をそのまま利用することができない．伝統的な統計学では選択バイアスを補正する方法として多重検定補正を用いてきたが，多数の仮説に対する多重検定補正は過度に保守的となり，有力な仮説が検出されずに見逃されてしまう．したがって，ビッグデータに基づいて生成された複数要因仮説の統計的信頼性を適切に評価するための情報数理的な方法論を整備することが必要であった．

２．研究の目的

本研究では，1)科学研究におけるビッグデータを用いて複合要因仮説を生成すること，2)ビッグデータ駆動型複合要因仮説の信頼性評価を行うこと，3)それらの成果を材料科学，生物科学，医療科学で実証することを目的とした

高次元データに対して複数要因の組み合わせを考えると，候補となる仮説の数が指数的に爆発して膨大なものとなってしまう．例えば，10,000 個の遺伝要因を計測したデータから，5 種類の遺伝要因に基づく複数要因仮説を考える場合，仮説の候補数は組み合わせ的に爆発し，10^{16} 個を上回る．莫大な仮説候補から最適な仮説を選択する場合，選択される可能性のない仮説候補を効率的に排除できるアプローチを確立することを目的とした．

従来の科学研究では，研究者の知識や経験に基づいて仮説が生成され，それを実験データに基づいて検証するという方式であった．データに基づいて仮説の生成と検証を行う場合，データに依存した仮説が生成されるため，仮説検定において「選択バイアス」が生じる．従来は多重検定補正の枠組によって選択バイアスを補正してきたが，補正が過度に保守的なために有力な仮説が見逃されてしまうリスクが大きい．近年注目されている「選択的仮説検定」と呼ばれるアプローチを用い，複数要因仮説の検証を行う方法を整備することを目的とした．

本研究で整備する情報数理的な方法論を材料科学，生物科学，医療科学の課題にて実証することを目的とした．それぞれの分野共同研究者と共同で実施し，各分野において科学的発見を行うことを目的とした．

３．研究の方法

目的 1 の複合要因仮説の生成に関しては，パターンマイニングとスパースモデリングの技術を統合させるアプローチに着目した．複合要因仮説は組み合わせ爆発が起こり，考慮すべき仮説の数が指数的に増加してしまうため，パターンマイニング分野で構築されたさまざまな計算技術を活用することでこの問題の解決を行った．パターンマイニング分野では，頻出パターンマイニングなど，いわゆる教師なし学習の分野におけるパターンの探索が主に研究されているが，予測モデリングなどの教師あり学習の枠組にそのまま利用するのは難しい．そこで，本研究ではスパースモデリングの概念を導入し，予測モデリングに不可欠なパターンをスパースモデリングの枠組で効率的に同定する方法を検討した．

目的 2 の複合要因仮説の評価に関しては，選択的推論と呼ばれる技術に着目した．選択的推論とは，仮説を選択するアルゴリズムの出力で条件づけたうえで統計的推測（統計的仮説検定や信頼区間推定）を行う枠組であり，2010 年代後半より，統計科学と機械学習において注目されつつあるものである．研究開始当初，通常のスパースモデリングに対する選択的推論の研究が行われつつあり，本研究では，その技術をパターンマイニングへ拡張する試みを行った．パターンマイニングの複雑なアルゴリズムで条件付けを行うには，新たな計算技術の開発が不可欠であり，本

研究では，問題の定式化と計算技術の開発を行った．

目的 3 の材料科学，生物科学，医療科学への適用は，それぞれの分野共同研究者と共同で行うこととした．材料分野の課題として，プロトン伝導体のポテンシャルエネルギーに大きく影響を与える原子配置パターンを同定する課題に取り組んだ．生物分野の課題として，吸光タンパク質として知られるロドプシンの吸光波長に影響を与えるアミノ酸の組み合わせ要因を同定する課題に取り組んだ．医療分野の課題として，心筋梗塞の疾病確率に影響を与える複数の一塩基多型（SNP）の組み合わせ要因を同定する課題に取り組んだ．

４．研究成果

本研究の実施により，3 つの目的を達成することができた．

1 つ目の目的である複合要因仮説の生成方法に関しては，セーフスクリーニングと呼ばれるスパースモデリングの技術をパターンマイニングに適用できるように拡張した．この方法はセーフパターンプルーニングと呼ばれ，スパースモデルで選択される複合要因仮説の枝刈りを効率的に行うことができるものである．本研究では，組み合わせ仮説のみでなく，グラフ構造データに基づく複合要因仮説や系列データに基づく複合要因仮説を生成する方法なども合わせて開発し，その成果は数多くの機械学習分野のトップジャーナルやトップカンファレンスに掲載された．

2 つ目の目的である複合要因仮説の評価方法に関しては，選択的推論をパターンマイニングに導入することで実現した．選択的推論では，条件付き推論を行うため，仮説の選択に関する制約条件を列挙する必要があるが，パターンマイニングに対する選択的推論では，考慮すべき制約条件の数が指数的に爆発してしまうという問題が生じた．本研究では，統計的推測を行ううえで，考慮すべき制約条件を効率的に探索する方法を確立し，考慮しなくてもよい制約条件を枝刈りによって効率的に排除することができるようなった．この方法は数多くの機械学習分野のトップジャーナルやトップカンファレンスに採択された．

3 つ目の目的である材料科学，生物科学，医療科学における実証では，それぞれの分野の共同研究者と共同研究を実施し，それらの成果は各分野のトップジャーナルに掲載された．

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　10件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Yoshida Tomoki、Takeuchi Ichiro、Karasuyama Masayuki	４．巻 NA
２．論文標題 Learning Interpretable Metric between Graphs	５．発行年 2019年
３．雑誌名 Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2019)	６．最初と最後の頁 NA
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3292500.3330845	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Umezu Yuta、Takeuchi Ichiro	４．巻 2
２．論文標題 Selective inference via marginal screening for high dimensional classification	５．発行年 2019年
３．雑誌名 Japanese Journal of Statistics and Data Science	６．最初と最後の頁 559～589
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42081-019-00058-8	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Ndiaye Y., Le T., Fercoq O., Salmon J., Takeuchi I.	４．巻 NA
２．論文標題 Safe Grid Search with Optimal Complexity	５．発行年 2019年
３．雑誌名 Proceedings of International Conference on Machine Learning (ICML2019)	６．最初と最後の頁 NA
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Karasuyama Masayuki、Inoue Keiichi、Nakamura Ryoko、Kandori Hideki、Takeuchi Ichiro	４．巻 8
２．論文標題 Understanding Colour Tuning Rules and Predicting Absorption Wavelengths of Microbial Rhodopsins by Data-Driven Machine-Learning Approach	５．発行年 2018年
３．雑誌名 Scientific Reports	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-018-33984-w	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Jalem Randy、Kanamori Kenta、Takeuchi Ichiro、Nakayama Masanobu、Yamasaki Hisatsugu、Saito Toshiya	4．巻 8
2．論文標題 Bayesian-Driven First-Principles Calculations for Accelerating Exploration of Fast Ion Conductors for Rechargeable Battery Application	5．発行年 2018年
3．雑誌名 Scientific Reports	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-018-23852-y	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1．著者名 Yonezu Tomohiro、Tamura Tomoyuki、Takeuchi Ichiro、Karasuyama Masayuki	4．巻 2
2．論文標題 Knowledge-transfer-based cost-effective search for interface structures: A case study on fcc-Al [110] tilt grain boundary	5．発行年 2018年
3．雑誌名 Physical Review Materials	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1103/PhysRevMaterials.2.113802	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1．著者名 K. Kanamori, K. Toyoura, J. Honda, K. Hattori, A. Seko, M. Karasuyama, K. Shitara, M. Shiga, A. Kuwabara, I. Takeuchi	4．巻 97
2．論文標題 Exploring a potential energy surface by machine learning for characterizing atomic transport	5．発行年 2018年
3．雑誌名 Physical Review B	6．最初と最後の頁 NA
掲載論文のDOI（デジタルオブジェクト識別子） 10.1103/PhysRevB.97.125124	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1．著者名 T. Tamura, M. Karasuyama, R. Kobayashi, R. Arakawa, Y. Shiihara, I. Takeuchi	4．巻 25-7
2．論文標題 Fast and Scalable Prediction of Local Energy at Grain Boundaries: Machine-learning based Modeling of First-principles Calculations	5．発行年 2017年
3．雑誌名 Modelling and Simulation in Materials Science and Engineering	6．最初と最後の頁 75003
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1．著者名 K. Aoki, H. Nakamura, H. Suzuki, K. Matsuo, K. Kataoka, T. Shimamura, K. Motomura, F. Ohka, S. Shiina, T. Yamamoto, Y. Nagata, T. Yoshizato, M. Mizoguchi, T. Abe, Y. Momii, Y. Muragaki, R. Watanabe, I. Ito, M. Sanada, H. Yajima, N. Morita, I. Takeuchi, S. Miyano, T. Wakabayashi, S. Ogawa, A. Natsume	4．巻 NA
2．論文標題 Prognostic relevance of genetic alterations in diffuse lower-grade gliomas	5．発行年 2017年
3．雑誌名 Neuro-Oncology	6．最初と最後の頁 NA
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/neuonc/nox132	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Y. Yasukochi, J. Sakuma, I. Takeuchi, K. Kato, M. Oguri, T. Fujimaki, H. Horibe, Y. Yamada	4．巻 NA
2．論文標題 Identification of CDC42BPG as a novel susceptibility locus for hyperuricemia in a Japanese population	5．発行年 2017年
3．雑誌名 Molecular Genetics and Genomics	6．最初と最後の頁 NA
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00438-017-1394-1	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計13件（うち招待講演　7件／うち国際学会　0件）

1．発表者名 竹内一郎
2．発表標題 Selective Inference による教師なし学習結果の信頼性評価
3．学会等名 統計学と機械学習の数理と展開（招待講演）
4．発表年 2019年

1．発表者名 竹内一郎
2．発表標題 データ駆動型人工知能のものづくりへの活用
3．学会等名 電子情報通信学会東海支部講演会（招待講演）
4．発表年 2019年

１．発表者名
竹内一郎

２．発表標題
データ駆動型科学のための統計的推論法

３．学会等名
情報理論とその応用シンポジウム（SITA2018）（招待講演）

４．発表年
2018年

１．発表者名
竹内一郎

２．発表標題
Selective Inference を用いた不均一データ分析のための統計的推論

３．学会等名
統計関連学会連合大会（招待講演）

４．発表年
2018年

１．発表者名
稲津佑，椙田大輔，豊浦和明，竹内一郎

２．発表標題
ガウス過程の導関数に基づく極小点の同定のための能動学習

３．学会等名
第21回情報論的学習理論ワークショップ（IBIS2018）

４．発表年
2018年

１．発表者名
井上茂乗，梅津佑太，竹内一郎

２．発表標題
そのクラスタ信用できますか? -クラスタ分割に対する統計的検証-

３．学会等名
第21回情報論的学習理論ワークショップ（IBIS2018）

４．発表年
2018年

１．発表者名
竹内一郎，金森研太，豊浦和明，本多淳也，服部，世古敦人，烏山昌幸，設楽一樹，志賀元紀，桑原彰秀

２．発表標題
機械学習による伝導性材料の物性値推定

３．学会等名
日本物理学会（招待講演）

４．発表年
2018年

１．発表者名
竹内一郎，佐久間拓人，西和弥，梅津佑太，岸本薫，烏山昌幸，梶岡慎輔，山崎修平，木村幸太郎，松本祥子，依田憲，福冨又三郎，設楽久志，小川宏人

２．発表標題
機械学習によるバイオロギングデータ分析

３．学会等名
日本生態学会（招待講演）

４．発表年
2018年

１．発表者名
竹内一郎

２．発表標題
データ駆動型の科学的発見とその材料科学への応用

３．学会等名
金属学会セミナー（招待講演）

４．発表年
2018年

１．発表者名
烏山昌幸，竹内一郎

２．発表標題
系列データからのクラス特異的代表パターン選出：分類モデルとMorse Complex によるアプローチ

３．学会等名
電子情報通信学会 第32回情報論的学習理論と機械学習研究会(IBISML)

４．発表年
2018年

1．発表者名 吉田知貴，竹内一郎，烏山昌幸
2．発表標題 マージン最大化距離学習におけるセーフスクリーニング
3．学会等名 電子情報通信学会 第31回情報論的学習理論と機械学習研究会(IBISML)
4．発表年 2017年

1．発表者名 井上茂乗，梅津佑太，坪田庄真，竹内一郎
2．発表標題 ヘテロジニアスなデータに対するクラスタリング後の推論
3．学会等名 電子情報通信学会 第31回情報論的学習理論と機械学習研究会(IBISML)
4．発表年 2017年

1．発表者名 米津智弘，田村友幸，小林亮，竹内一郎，烏山昌幸
2．発表標題 コスト考慮型ベイズ最適化による複数目的関数最適化とその材料分野への応用
3．学会等名 電子情報通信学会 第29回情報論的学習理論と機械学習研究会(IBISML)
4．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	二宮　嘉行 (Ninomiya Yoshiyuki) (50343330)	統計数理研究所・数理・推論研究系・教授 (62603)	

6．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	豊浦　和明 (Toyoura Kazuaki) (60590172)	京都大学・工学研究科・准教授 (14301)	
研究分担者	安河内　彦輝 (Yasukochi Yoshiki) (60624525)	三重大学・地域イノベーション推進機構・助教 (14101)	
研究分担者	井上　圭一 (Inoue Keiichi) (90467001)	東京大学・物性研究所・准教授 (12601)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関