戦略的創造研究推進事業　CREST
研究領域「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」
研究課題「離散構造統計学の創出と癌科学への展開」

研究終了報告書

研究期間　2015年10月～2021年3月

研究代表者：津田宏治
（国立大学法人東京大学大学院新領域創成科学研究科　教授）

§1 研究実施の概要

(1)実施概要

本課題では、ビッグデータ時代を切り開く新たな統計手法の開発と、その医学・生物学分野での応用を目指してスタートした。特に、小児がんの一種である神経芽腫の原因究明をターゲットに研究を進めてきた。本課題で得られた成果は、以下の項目にまとめられる。

選択的推論手法の開拓 (竹内 G、津田 G): 現在の実験科学においては、実験機器の進化によって説明変数の数が飛躍的に増加している。また、組み合わせ要因を考えると、考慮すべき仮説の数は、天文学的な数に達する。このような多数の仮説が存在する状況で、有効な検定値(P 値)を得るため、近年、選択的推論という新たな統計手法が提案されている。選択的推論では、多数の仮説を少数に絞り込む選択操作を定義した上で、選択操作に条件づけた P 値を計算するため、仮説の数が多い場合でも、十分な検定力を発揮することができる。しかし、画像分割、組み合わせ要因に対する特徴選択、クラスタリングなど、複雑な選択操作が与えられたとき、それに条件づけた P 値を計算するアルゴリズムを導出することはトリビアルではない。本課題では、いくつかの重要な選択操作に関し、選択的推論を行う効率的なアルゴリズムを提案した。まず、組み合わせ要因を考慮して回帰分析を行う、スパース高次交互作用モデルに対する選択的推論法を開発し、ICML 2017 において論文発表を行った。次に、グラフカットに基づく画像分割に関する選択的推論法を開発し、CVPR2020 において論文発表を行った。また、K-means クラスタリングを行った後、クラスタ間で分布が異なる変数を選択する問題(post clustering inference)に関して、選択的推論法を開発し、がん細胞の一細胞遺伝子発現データに対して適用した。過去の論文で公開されているデータに対して適用することで、過去の結果において統計的有意性が誇大に主張されている可能性を指摘した。

LAMP 法のさらなる発展と医療応用(瀬々G、津田 G): LAMP 法とは、瀬々、津田らによって提案された、データから組み合わせ要因を発見し、適切に多重検定補正が行われた P 値を計算する手法である。LAMP は原理的には様々な統計検定に適用できるが、これまでは、医療応用で重要な生存時間解析に適用できる手法は開発されてこなかった。本課題では、LAMP を生存時間解析に適用できるように拡張し、Survival LAMP と名付けた。この手法を 526 人の乳がん患者の細胞から得られた 14688 遺伝子の発現量に適用した結果、生存曲線に統計的有意に関連する遺伝子の組み合わせを発見することができた。また、国立循環器病研究センターと連携し、血圧、γ-GPT、中性脂肪などの項目を含む2万人以上の定期健康検診のデータに LAMP を適用することで、健康データに対する新たなアプローチを試みた。その結果、急性心筋梗塞の発症に統計的に有意に関連する定期検診の検査項目の組み合わせを発見することができた。今後、人間ドックでの使用など実用化が期待できる。

一細胞解析による神経芽腫の原因究明(門松 G):神経芽腫は多くの子供が罹患する小児がんであり、その原因は解明されていない。特に、約20%の患者が、自然に治癒する現象(自然退縮)に関しても、原因遺伝子は知られておらず、自然退縮のメカニズムを解明することで、治療法開発につなげられると考えられる。本課題では、マウスの一細胞 RNA-seq データの詳細な解析と、検証実験により神経芽腫の発生に関与する転写因子を発見した。

情報幾何に基づく統計検定手法の開発(山田 G): 情報幾何を基礎とする数学的な立場から検出力の高い新たな統計検定手法の研究を行い、MAX-test の多重検定補正法や、形状データに関する検定法の開発などにつながった。

ソフトウエアの開発(津田 G、瀬々G): LAMP を大規模データに対応適用するため、スーパーコンピュータ用ソフトウエア MP-LAMP を開発した。数千個の CPU を並列に使用することが可能である。

(2)顕著な成果

＜優れた基礎研究としての成果＞
1． 複数要因の組み合わせに関する選択的推論法の開発
概要:
選択的推論においては，仮説を選択するアルゴリズムが線形不等式制約の集合で表現されれば，その仮説が選択されたという条件に基づく検定統計量の条件付分布を計算することができる．本研究では，まず，パターンマイニング分野で開発された分岐限定法に基づく技術を用いることで，複数要因の組み合わせ効果に関する仮説を線形不等式制約によって表現できることを示した．さらに，膨大な数にのぼる線形不等式制約のうち，検定統計量の条件付帰無分布に影響を与えないものを予め同定するためのアルゴリズムを開発した．これらの成果を組み合わせることにより，複数要因の組み合わせ効果に関する仮説の生成と検証を Selective Inference の枠組で実施できることとなった．本研究の成果は機械学習分野の最難関国際会議の一つである ICML2017 に採択され，発表した．

2． 生存時間解析に適用可能な LAMP 法の開発
概要:
生存時間解析（Survival Analysis）は、癌研究などの臨床研究で重要な役割を果たす。我々が開発した組み合わせ要因発見・検定法 LAMP (Terada et al., PNAS 2013)においては、理論的に適用できるデータの種類に制限があり、生存時間解析に用いることはできなかった。本研究では、生存時間解析で一般的に用いられる log-rank test に関して P 値の下限を導出し、生存時間解析が可能なアルゴリズム Survival LAMP を開発した。TCGA データベースに含まれる乳房浸潤癌細胞の発現プロファイルに対して適用したところ、5種類の遺伝子の組み合わせを発見し、統計的に有意であることを確認した。

3．画像セグメンテーションにおける選択的推論手法の開発
概要:
本研究ではグラフカットによる病理画像セグメンテーションの統計的信頼性を評価するための選択的推論法を開発した．セグメンテーションのような教師なし学習結果の統計的信頼性を評価するのは困難であるが，選択的推論を導入することでバイアスのない統計的有意性評価を始めて可能とした．本論文ではグラフカットによる選択イベントが線形不等式と二次不等式の集合で表現できることを活用した．

＜科学技術イノベーションに大きく寄与する成果＞
1．LAMP 法による定期検診データの解析
概要:
健康診断後 5 年間に急性心筋梗塞を発症した患者に統計的に有意に関連する定期検診の項目の組み合わせを検出する解析を行った。利用したデータは、2011 年度の定期検診のデータであり、解析に利用した人数は、男性 19,384 名、女性 8,974 名である。MP-LAMP を用いて解析した結果、男性のデータでは、急性心筋梗塞の発症に有意に関連する健診項目の組み合わせ 1,448 個を検出した。今後、単体の検査項目だけに着目するのではなく、複数の検査項目の組み合わせを考慮することで、急性心筋梗塞の発症リスクが高い条件を発見することが期待できる。

2．LAMP を用いた震災被災者の PTSD データの解析
概要:

• 東日本大震災の被災地での PTSD 症状予後に関するコホート調査情報に対して、MP-LAMP を実行することで、有意なリスク因子の組み合わせについて網羅的な解析を行った。その結果、最も P 値の小さい組み合わせは「少ない休息」、「運動不足」、「非雇用」、「準備せずに避難」（P=2.2×10e-4)であり、組み合わせを考慮しない場合に比べて詳細な結果が得られた。

3．多重検定ソフトウエアの開発と公開

概要：

本課題で開発した統計検定手法を普及させるため、LAMPLINK、MP-LAMP という二種類のソフトウエアを公開した。 LAMPLINK は、DNA の変異を含む遺伝データ解析のためのパッケージであり。遺伝学分野で著名なソフトウエアである PLINK と同様の入力ファイルを使用することができるため、PLINK を使用したことがあれば、簡単に使用できる。MP-LAMP は、数千個の CPU を用いた並列処理のパッケージであり、LAMP 法を大規模なデータに適用する際には必須のパッケージである。これまでに京コンピュータや、AWS などのスーパーコンピュータシステムで動作を確認している。

＜代表的な論文＞

1．S. Suzumura, K. Nakagawa, Y. Umezu, K. Tsuda, I. Takeuchi, Selective Inference for Sparse High-Order Interaction Models, Proceedings of the 34th International Conference on Machine Learning (ICML 2017), pp. 3338-3347, 2017.

概要：

統計的に有意な高次相互作用を見つけ出すことは重要ではあるが、困難なタスクである。本論文では、スパース高次相互作用に対する選択的推論手法を提案する。我々の主たる貢献は、線形モデルに対して提案された選択的推論の枠組みを高次相互作用モデルに拡張することにある。提案アルゴリズムの有効性を確かめるため、HIV 薬剤耐性の予測問題への応用を行う。

2. R.T. Relator, A. Terada, J. Sese, Identifying statistically significant combinatorial markers for survival analysis, BMC Medical Genomics, 11(Suppl 2), 31, 2018.

概要：

我々は、生存時間解析において、生存に関わる組み合わせ要因を発見し検定する問題を取り扱う。ここでは、最近提案された LAMP 法を log-rank 検定に拡張することで、この課題を達成する。癌からの遺伝子組み合わせの発見問題に本手法を適用して有効性を確かめる。

3. K. Tanizaki, N. Hashimoto, Y. Inatsu, H. Hontani, I. Takeuchi. Computing Valid P-values for Image Segmentation by Selective Inference. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2020), 2020.

概要：

画像分割は、コンピュータビジョンにおける代表的なタスクである。本論文では、選択的推論という枠組みに基づいて、画像分割における統計検定を行う。グラフカットと、閾値ベースという二つの分割アルゴリズムに対して、アルゴリズムを開発し、医療画像への応用を通して有効性を確かめる。

§２ 研究実施体制

(1)研究チームの体制について

1. 津田グループ
研究代表者: 津田 宏治 (東京大学・大学院新領域創成科学研究科・教授)
研究項目:離散構造統計学の創出・普及

2. 門松グループ
主たる共同研究者:門松 健治 (名古屋大学・大学院医学系研究科・教授)
研究項目:癌検体の収集、実験データの取得および介入実験

3. 瀬々グループ
主たる共同研究者:瀬々 潤 (株式会社ヒューマノーム研究所・代表取締役社長)
研究項目:統計的検定手法構築、高速化、大規模化及び手法の普及

4. 竹内グループ
主たる共同研究者:竹内 一郎 (名古屋工業大学・大学院工学研究科・教授)
研究項目:網羅的遺伝情報の複合要因探索アルゴリズム構築・ソフトウェア実装・癌科学における実証

5. 山田グループ
主たる共同研究者:山田 亮 (京都大学・大学院医学研究科・教授)
研究項目:離散構造統計学の遺伝疫学・コホートスタディへの展開

(2)国内外の研究者や産業界等との連携によるネットワーク形成の状況について

現在、LAMP を始めとする統計的パターンマイニングに関する国際的なコミュニティが構成されており、アクティブに活動している。主要なメンバーは、Geoff Webb (Monash University), Wilhelmiina Hamalainen (Aalto University), Karsten Borgwardt (ETH Zurich)などである。2015 年に、本課題の特定課題調査の一環として、東京で Tokyo Workshop on Statistically Sound Data Mining を開催した。2017 年 6 月には、UC Riverside において開催された International Conference on Multiple Comparison Procedures (MCP)において、招待セッション「Pattern Mining under Multiplicity Control」を開催した。演題は以下のとおりである。

Statistical Pattern Mining: An Overview (Invited Talk)
Koji Tsuda*, University of Tokyo, Japan

Selective Inference for Predictive Pattern Mining
Ichiro Takeuchi*, Shinya Suzumura, Yuta Umezu, Koji Tsuda, Nagoya Institute of Technology, Japan

Accounting for a Categorical Covariate in Significant Pattern Mining
Llinares Lopez Felipe*, Laetitia Papaxanthos, Dean Bodenham, Damian Roqueiro, Karsten Borgwardt, ETH Zurich, Switzerland

Controlling Familywise Error When Rejecting at Most One Null Hypothesis Each From a Sequence of Sub-Families of Null Hypotheses
Geoff Webb*, Mark van der Laan, Monash University, Australia

また、このセッションの他にも、小宮山純平(東大)、杉山麿人(情報研)、山田誠(理研 AIP)による離

散構造統計学関連の発表が行われた。離散構造統計学の成果はデータマイニングや機械学習系の学会で発表されることが多かったが、今回、数理統計学の学会で、まとまった発表をすることができ、統計学者の認知度を高めることができた。最近でも、KDD2019 において、Hypothesis Testing and Statistically-sound Pattern Mining と題したチュートリアルが開催されるなど、LAMP 法によって種がまかれた離散構造統計学の分野は多くの研究者を巻き込んで発展している。

2017 年 2 月 16 日(木), 17 日(金)に名古屋工業大学において，シンポジウム「ビッグデータ利活用のための基盤構築とその応用」を開催した．本シンポジウムでは，数理統計や機械学習の若手研究者を中心に 16 件の講演を行った．講演者の専門分野をもとに，近年のデータ解析に重要な手法や応用についてご講演いただいた(位相データ解析，ロバスト推定，関数データ解析，介護保険データへの応用など)．また，参加者数は，講演者を含め 40 名以上(延べ約 80 名)に上った．

産業界との連携に関しては、瀬々によって創業された(株)ヒューマノーム研究所によって、医療機関・大学病院等のリアルデータを用いた異分野共同による解析が開始されている。医療連携に関しては、東北大学、国立循環器病研究センターの複数の研究者と連携し研究成果を創出した。

竹内 G では，選択的推論に関する共同研究を Université de Montpellier(フランス)の Joseph Salmon 教授と行った．同教授の教え子であった Ndiaye Eugene 氏とは，現在，日本で博士研究員として共同研究を行っている．また，名古屋大学医学研究科病理教室と共同研究を行い，医療診断結果の信頼性に関する共著論文を発表した．