TSUBAME 共同利用　令和4年度　学術利用　成果報告書

臨床情報統合データベースの機械学習解析
Machine Learning for Integrated Database of Clinical Information

浅井　聰
Satoshi Asai

日本大学　医学部　生体機能医学系　薬理学分野
Department of Pharmacology, School of Medicine, Nihon University
https://www.med.nihon-u.ac.jp/department/pharmacology/

臨床情報統合データベースに含まれる薬物投与・過去の診断や検査結果の履歴から疾患の発症・合併症進行リスクを予測するベイズモデルを構築しようとした。この目的のため、ベイズモデルにおいて少数のリスク因子を非線形効果まで含めて精密に評価するために、アニールド重点標本可逆ジャンプリーマン多様体上ハミルトンモンテカルロ法の大規模実装を開発した。効率化の工夫により従来の可逆ジャンプモンテカルロ法と異なり、大規模データのベイズマルチカーネル学習に実用可能なものが得られた。さらにこれまで開発した手法を実際の臨床データに適用し、既存の医療統計手法では検出できなかったバイアス効果を明らかにする等の結果を得た。

We developed an efficient large-scale implementation of Bayesian multiple-kernel learning based on annealed-importance-sampling-Riemannian-manifold-Hamiltonian Monte Carlo, with the goal of estimating risks of disease onset and progression based on clinical information recorded in a large database of electronic medical records. Developing and applying different techniques for efficient computation, we obtained an implementation of the algorithm that can deal with a Bayesian multiple-kernel model with large numbers of kernels and samples, unlike previously reported reversible-jump algorithms.

Keywords: electronic medical records, pharmacoepidemiology, Bayesian analysis, multiple-kernel learning, Monte-Carlo method

背景と目的

本研究の背景及び目的は昨年度の我々の同名の課題と概ね同じであり、まずこれを以下に引用する。（**引用開始**）近年、電子カルテデータなどを含む大規模な医療データベースから薬剤の効果・疾患リスクなどの医学的知見を機械学習・人工知能技術を用いて抽出する試みが注目を集めている。実際に、世界各国の大学病院や地域中核病院の電子カルテデータを用いた解析が行われはじめており[1,2]、さらに適切な倫理的な枠組みのもと、電子カルテデータに患者から採取した遺伝·生化学的測定データを組み合わせ、これを患者個別医療に役立てようとする試みがはじまっている。

しかしながら、近年の機械学習·人工知能技術の発展にもかかわらず、上記のような大規模医療データから有益な医学的知見を抽出するためには解決すべき技術的課題が複数存在する。例えば、医療データベースの解析では薬剤投与・検査値推移・疾患発症など数千以上の予測変数から疾患進行を予測しようとするが、実際に疾患進行に影響する変数は少数であり、変数選択を効率的に行う必要がある。また臨床医が判断する重症度のような潜在変数の情報をアルゴリズムによって推定できることも必要だ。これらを変数間の非線形関係も考慮しながら行うことは，潜在変数付きノンパラメトリックベイズモデルでスパースな変数選択を行う問題に帰着される(次ページ図参照)。上記のような潜在変数付きのノンパラメトリックベイズの枠組みにおいて、医療データベースのような大規模データ上でスパースかつ効率的な推定に成功した研究はまだ存在しない。先行研究のほとんどで潜在変数のモデリングは巧妙に避けられている。実際、医療データに限らずノンパラメトリックベイズモデルにおける潜在変数の推定やスパース変数選択は、機械学習一般の問題として依然として困難な問題の一つであり、効率的な解法が模索されている[3]。申請者のグループは、下にも述べるように、この問題を解決しうる新奇なアルゴリズムを得たので、本課題において TSUBAME を利用した大規模実装を行いこ

の手法の有用性を示そうとする。**（引用終了）**

本年度では上記に加えて、これまでに開発した機械学習アルゴリズムを具体的にデータベースの解析に適用していくという目的を設定した。そして、具体的に解析を行ったところ、従来の医療統計の手法の限界を超えて新たな知見を得られたため、その内容を以下にまとめる。また、新奇アルゴリズムの開発についても一定の成果が得られ、論文報告の準備中であり、内容を以下にまとめる。

概要

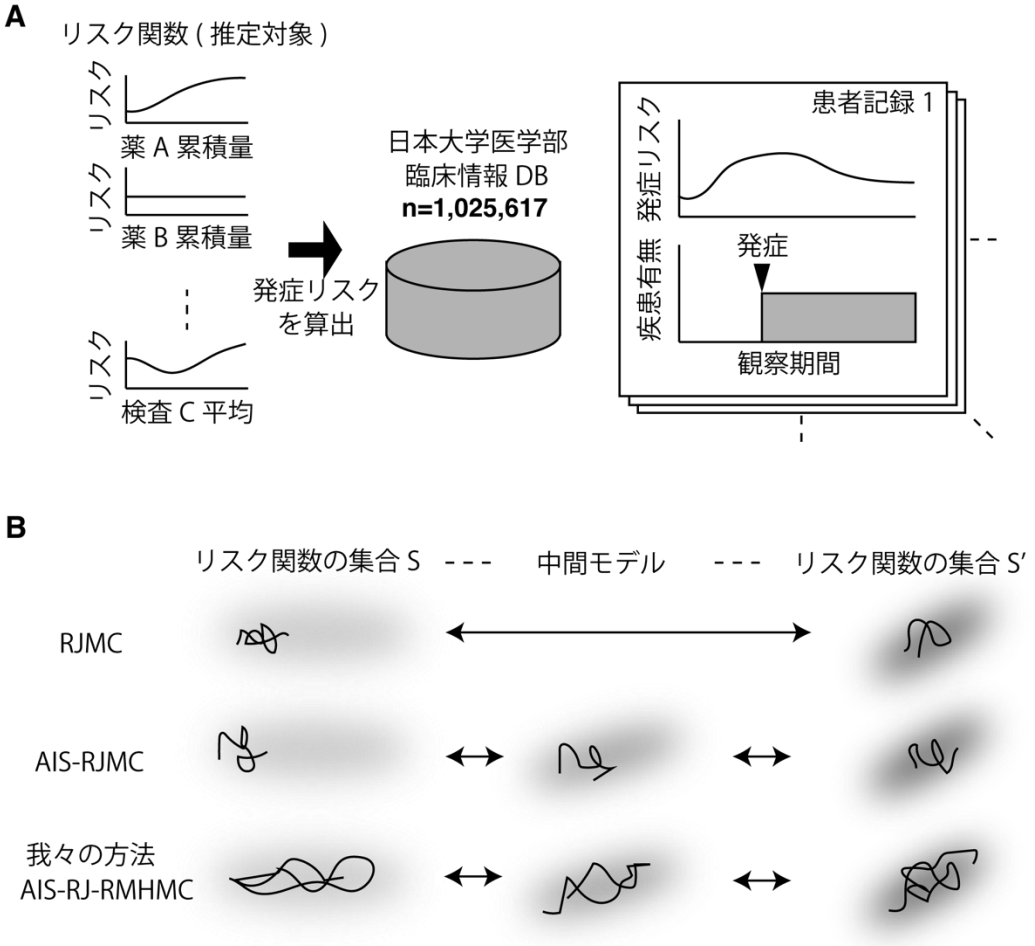
臨床情報統合データベースから医学的知見を引き出すための新奇アルゴリズムの大規模実装が部分的に完成した。昨年度から引き続きガウス過程に基づいたノンパラメトリックベイズの大規模実装を可能にするべく、高効率のモンテカルロ法を開発した。リーマン多様体上のハミルトンモンテカルロ法とアニールド重点サンプリングとよばれる手法を用いつつ独自の実装上の効率化を施すことで高効率のベイズマルチカーネル法の事後過程サンプリングアルゴリズムの開発に成功した。

また、昨年度開発した、非ベイズ設定におけるマルチカーネル法を実際にデータベースに適用し、新たな医学的知見を得るとともに、開発中のアルゴリズムを実際にデータベースに適用するためのインターフェースを開発した。具体的には、近年論争の題材となっていた、ベンゾジアゼピン系薬物・抗コリン作用を持つ薬物の長期投与による認知症リスクに関する従来の研究結果に多分にバイアスが含まれている可能性が高いことを機械学習手法を用いて明らかにした。

結果および考察

<u>ベイズマルチカーネル法の大規模実装</u>

実装すべきアルゴリズムは下図 A のように、薬物投与量に対して疾患・合併症発症などのような予測すべき対象のリスク関数を設定し、それらを統計的に推定することである。このとき、疾患発症の予測に役立つ変数の数は、比較的少数と考えられ、これらの少数の変数が何であるかを見抜きたい。また、疾患リスクは薬物投与量などに対して非線形に変化する



ことが通常であり、非線形推定の手法を用いたい。将来的に潜在変数を導入することからベイズの枠組みでアルゴリズムを構築することを考慮すると、ベイズマルチカーネル法が適しているという結論になった。ベイズマルチカーネル法はこれまでも理論的には論じられてきたが[4]、技術的な困難から実用上はあまり用いられてこなかった。議論は省略するものの、いくつかのアプローチの中で我々はノンパラメトリックベイズの枠組みで変数選択をする上で根本的に問題を解決する可能性が最も高いと思われる可逆ジャンプモンテカルロ法を用いた。この方法は 1995 年に Green らによって提唱されて以来[5]、用いる変数・カーネルの集合を変更する際の可逆ジャンプの採択率の悪さが問題とされ、様々な改良案が提案されていた。我々は、図 B に示すように、あるリスク関数の集合から、別のリスク関数の集合にジャンプする際に、それらのリスク関数集合に対応するモデルを人工的に内挿した中間モデルを多数用意しこれらの間を可逆的にジャンプすると、計算時間を増やせば増やすほど理想的な採択率に近づくというアルゴリズムを定式化したが、調査の結果これはアニールド重点標本可逆ジャンプモンテカルロ法(AIS-RJMC)として 2013 年に一般的･抽象的な形で Karagiannis と Andrieu により提案されている方法[6]であった。しかしながら、この方法は実応用上での高性能にまだ結びついていないためか、あまり応用例が報告されていなかった。なぜ応用上の高性能に結びついていないかは明らかで、理想的な採択率の実現のためには可逆ジャンプの一回一回においてモンテカルロ法を収束させたまま一つのモデルから別のモデルにゆっくりと移行させる必要があったが、通常の AIS-RJMC の実装ではそのようなことが容易でなかったからである(図 B)。しかしながら我々は、ベイズマルチカーネルの問題設定において応用上十分な精度でこれを達成することに成功した。具体的には、モンテカルロ法としてスケール不変性を持つエネルギーの修正ヘッセ行列(ヘッセ行列の負固有値を正に固有値に置き換えたもの)を計量とするリーマン多様体上のハミルトンモンテカルロ法を組み込んだ AIS-RJ-RMHMC を設計した。リーマン多様体上のハミルトンモンテカルロ法(RMHMC)は Girolami らにより 2011 年に提案されているが、その実行のためには、各モンテカルロ試行で毎回修正ヘッセ行列の全ての 3 回偏微分や逆行列を計算する必要があり、計算コスト上の難点があった。しかしながら我々は、修正ヘッセ行列の 3 回微分をもとのヘッセ行列の微分値と特異値分解を用いて高速に計算しつつ、前状態のヘッセ行列の特異値分解の結果を出発点として現在のヘッセ行列の特異値分解に 2 次収束する Jacobi 法をベースとするアルゴリズムを用いて計算することでこの計算量を大幅に低減させ、実用に足るアルゴリズムを得た。

このアルゴリズムを、機械学習の標準ベンチマークデータである UCI データセットに適用した結果、及び以前我々が用いた日本大学医学部における認知症リスク解析のためのデータセットに適用した結果と合わせて論文公表を準備中である。

非ベイズマルチカーネル法の大規模実装の日本大学医学部臨床情報統合データベースへの応用

令和3年度から令和4年度初頭に我々はベイズマルチカーネル法の実装に先立って非ベイズのマルチカーネル法の TSUBAME3.0 における複数 GPU を利用した大規模実装を構築していた。そこで、日本大学医学部倫理委員会の承認のもと、この大規模実装の臨床情報統合データベースへの適用を開始した。特に令和 4 年度後半は、まず認知症リスクの解析に注力した。認知症のリスクは多数知られているが、特に長期投与されている薬物によるリスクが論争を引き起こしていた。従来法による医療統計解析によりベンゾジアゼピン系薬物や抗コリン作用を持つ薬物の長期投与が認知症リスクを高めるとする複数の研究結果が報告されていた一方、これは統計解析のデザインに起因する逆因果バイアス・交絡バイアスではないかという疑いを唱える論文も発表されていた。そこで、我々は逆因果やその他の交絡を表す予測変数・カーネルをこれらの薬物の効果を表す予測変数・カーネルと同時に用いて統計解析を行ったところ、逆因果・その他の交絡の影響が明確に示される結果となった。そしてこれらのバイアスの調整

下では(少なくとも我々のデータセットからは)これらの薬物の長期投与による認知症発症リスクの上昇は有意でないという結果を得た。ベンゾジアゼピン系薬物の効果の解析結果については論文投稿後、出版社から小修正後に受理する連絡を受けており、抗コリン作用を持つ薬物の影響については論文投稿準備中である。

まとめ、今後の課題

完成した大規模実装によって効率的な変数選択を行いながら疾患リスクや薬物効果を判定できるようになった。しかしながら、潜在変数を組み込んだモデルのための実装はまだ行っていない。臨床情報統合データベースのデータのような時系列データの潜在変数の推定のためには、祖先サンプリング付き粒子ギブス法という確立した手法があり、これを上記の AIS-RJ-RMHMC に組み込むことで達成できると考えられるが、データが大規模であるだけにさらなる入念な調整が必要と思われる。令和 5 年度はこの点においてアルゴリズムを向上させつつ大規模応用の応用範囲を広げていくことを展望としている。

参考文献

[1]S. Lee and H.-S. Kim J Lipid, Atheroscler. (2021) 10(3):282-290
[2]M. Chowdhury et al., Front. Psychiatry (2021) 738466
[3]M. Gönen, Proceedings of ICML (2012)
[4]T. Suzuki, CoLT 2012, pp.8.1-8.20
[5]P. Green, Biometrika (1995) 82(4):711-732
[6]G. Karagiannis & C. Andrieu, J. Comp. Graph. Stat. (2013) 22(3): 623-648
[7]M. Girolami & B. Calderhead, J. R. Stat. Soc. B (2011) 73(2): 123-214