TSUBAME 共同利用　令和3年度　学術利用　成果報告書

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

利用課題責任者　浅井　聰
Satoshi Asai

所属　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://nu-pharmacology.com

本課題は医療データベースから生活習慣病に関する知見を抽出するためにノンパラメトリックベイズ法の効率的な大規模アルゴリムを実装することを目的とする。特にノンパラメトリックベイズ法においてスパース性と潜在変数を扱えるモンテカルロアルゴリズムを構築し、データの背後にある潜在的医学生物学的メカニズムを抽出しつつこれに最も影響を与えるリスク要因を明らかにしようとする。この目的で、申請者のグループが開発した新奇な手法を GPU アクセラレーターを利用して大規模データベースに適用可能にする。本年度は既存のアルゴリズムを GPU にオフロードし高速化することに成功し、これを改変することで提案法のコードを開発した。

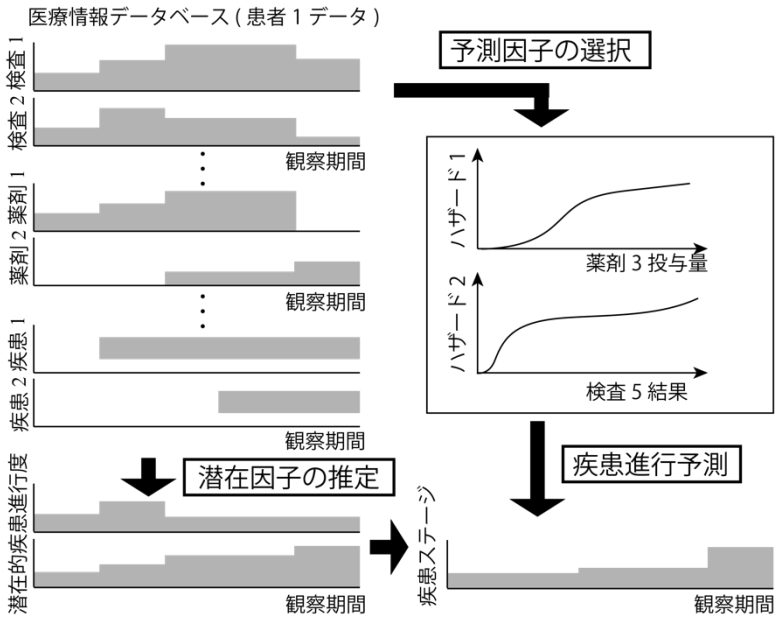
We aim at extracting information about biophysiological mechanisms and key factors for the onset and progression of metabolic diseases from a large volume of electric medical records. For this purpose, we develop an efficient, sparse nonparametric Bayesian algorithm with latent variables based on Monte-Carlo sampling. In the fiscal year 2021, we succeeded to develop a code that drastically accelerates a conventional algorithm by using GPU accelerators, and then, we developed a code for our new method by modifying this code.

Keywords: electric medical record, nonparametric Bayes, Monte-Carlo algorithm, GPU computing, metabolic diseases

背景と目的

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と一定の理論的保証を得たので、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。本年度はこの実装に一定の進展があったので報告する。



図： 医療情報データベースに基づく疾患進行予測と予測因子・潜在因子の推定

概要

上記「背景と目的」に述べたように、大規模医療データベースから医学的知見を引き出すには潜在変数付きのノンパラメトリックベイズモデルにおいて効率的な変数のスパース選択を可能にする必要がある。申請者のグループではこれを可能にするモンテカルロアルゴリズムを設計し、その理論的性能保証を得た。そこで実際に大規模医療データベースに適用するための大規模実装を開発し、その性能を実証しようとする。特にTSUBAMEの複数GPU環境を利用することで大規模データでも現実的な時間内に推論が可能であることを示し，さらにアルゴリズム面でも従来法との比較において提案手法の優位性を示す．その後，倫理委員会の承認を経た上で実際の匿名化データにアルゴリズムを適用し医学的有用性を示す．

結果および考察

提案法を実装する準備として，非ベイズの設定におけるマルチカーネル学習の従来法をTSUBAME3.0 上で大規模実装し，動作確認を得た。電子カルテデータで想定される百万サンプル×数百～数千予測変数での計算をC++とOpenaccにより，全て GPU にオフロードすることで高速計算が可能となった。この従来法の実装は申請者らの提案法との比較において役立つとともに、この実装の改変によって提案法を実装していくための実験コードとして役立つと期待される。実際、この結果から申請者らは提案法の場合の大規模実装の見通しを得るとともに、実装されたコードを改変しながら提案法のコードの開発も進めることができた。

まとめ、今後の課題

TSUBAMEでの課題に取り組み始めて4ヶ月であり課題解決にさらに時間を要するが、もう数ヶ月で提案法の実装が完了する見通しである。完了後はベンチマーク用データを用いてアルゴリズムの性能を従来法と比較し、提案法の有用性を示し論文発表する。また、倫理委員会の承認を経て実際の医療データにアルゴリズムを適用していく。

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)