様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ４ 月 ２９ 日現在

機関番号：１４４０１
研究種目：挑戦的萌芽研究
研究期間：2014～2015
課題番号：２６５４０１１６

研究課題名（和文）モデルマイニング：超高次元大規模データからの局所モデル探索列挙手法の探求

研究課題名（英文）Model Mining: Exploration of search and enumeration methods of local models from super-high dimensional data

研究代表者
鷲尾　隆（Washio, Takashi）

大阪大学・産業科学研究所・教授

研究者番号：００１９２８１５

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：　超高次元大規模データから各サブプロセスを表すモデルを高速探索するモデルマイニング原理を探求し、医療等への実験適用を通じたアルゴリズム検討を目的とした。その結果、大規模データから高速、高精度なモデルマイングが可能なランダムサブサンプリングとアンサンブルモデリングの原理を確立し、それを実装する半空間データ質量や類似性尺度の手法を得た。またそれらを医療分野に適用し、臨床患者データから新しい心疾患生起メカニズムモデルを発掘することに成功した。以上の成果を、機械学習の世界トップ論文誌であるMachine Learningやデータマイニングの世界トップ国際会議であるICDM及び医学主要論文誌に発表した。

研究成果の概要（英文）：This study aimed at the exploration of model mining principles, which enable fast search of candidate models representing sub-processes embedded in super-high dimensional and large scale data, and their implementations into some algorithms for applying to experimental problems including medical fields. We established novel principles of random sub-sampling and ensemble modeling for fast and accurate model mining from the large scale data, and developed the methods of half-space mass and and mass based similarity measures by implementing the principles. Finally, by applying these methods to heart disease data in medicine, we succeeded to mine a model of a occurrence mechanism of the heart disease. These outcomes have been presented in Machine Learning:the world top journal of machine learning, ICDM: the world top international of data mining and a major medical journal.

研究分野：Data Mining

キーワード：データマイニング　機械学習　ビッグデータ　モデリング　高次元データ　サンプリング　アンサンブル

１．研究開始当初の背景

ビッグデータの普及で、膨大なサブプロセスの条件・状態組合せを含む大規模対象系のモデリング需要が増している。しかし、条件・状態組合せ爆発によりビッグデータでさえ系全体のモデリングには圧倒的に不足で、有効なデータマイニング・機械学習手法はない。

筆者等は自らの医療・経済分野での実践研究経験から、大量の未知サブプロセスの条件・状態組合せを含む超高次元大規模対象系のモデリングには、条件・状態組合せ爆発によりビッグデータでも圧倒的に足りず、しかも各サブプロセスを反映する観測変数・事例が無関係な大量変数・事例に埋もれ、対象系の全体はおろか局所モデルすら一意同定が困難なことを認識した。この克服には、各未知サブプロセスに対応する関連観測変数・事例の絞込みによる組合せ爆発回避と多数の局所モデル候補導出を同時に行うマイニング原理が必要である。

データマイニングには、筆者等の手法を含め変数部分空間とその事例クラスタを同時探索する部分空間クラスタリング技術があるが、特定モデルを構成する変数部分空間や事例集合の発掘ではない。機械学習でも筆者等の手法を含め、正則化による主要サブプロセスの変数選択とモデリング、事例選択によるモデリング、混合モデルやディープラーニングによる複雑モデリングがあるが、何れも対象系の全体モデルを一意同定する技術である。

研究開始当初、超高次元大規模データから各サブプロセスモデル候補の高速探索列挙を行うモデルマイニングの着想は世界的にもなかった。

２．研究の目的

本研究では、数千次元を超える超高次元大規模データから各サブプロセスを表す変数と事例、モデルの候補組を高速探索列挙するモデルマイニング原理を探求すること、さらに医療等への実験適用を通じアルゴリズムを検討することを目的とした。筆者等は、自らの部分空間クラスタリング、最適化変数・事例選択、探索列挙等の研究成果と医療・経済等の実践的研究経験からこの着想を得た。本探究により、世界を先導する次世代データマイニング・機械学習の重要分野を拓くことを目指した。

より具体的には、筆者等の既存成果に同じく探索列挙の研究成果を加え、数千次元を超える超高次元大規模データに埋もれる各局所モデルとその変数部分集合、事例部分集合の候補を高速探索列挙するモデルマイニング手法を探求した。これらはさらに、

（１）データから蓋然性の高い候補を見出す統計的・情報論的原理の構築
（２）超高次元大規模データから統計的・情報論的原理により候補を高速探索する手法の確立
（３）医療などの実ビッグデータに基づく効率的アルゴリズムの検討
（４）原理・手法・アルゴリズムの実例題検証

の研究項目に別れ、これらの遂行により世界を先導する次世代データマイニング・機械学習の方法論を提示することを目指した。

３．研究の方法

２年間で瞬発に成果を出すことを目指し、平成２６年度は(1) データから蓋然性の高い候補を見出す統計的・情報論的原理の確立、(2) データから統計的・情報論的原理により候補を高速探索列挙する手法の構築に取り組んだ。２７年度は(1)(2)を継続しつつ(3)及び(４)原理・手法・アルゴリズムの実例題検証を重点とし、モデルマイニングの基礎の確立に取り組んだ。

実施項目(1)では、蓋然性の高い変数部分集合,事例部分集合,モデルを見出す原理を、モデルの(Ⅰ)単純さ、(Ⅱ)精度、(Ⅲ)外れ事例の３要素を考慮してした。従来の統計・機械学習では、データ分布とモデルによる事後分布の KL-Divergence に相当する AIC から発展した ABIC や GIC などの情報量基準やベイズ周辺尤度に基づく BIC、最小記述長等、(Ⅰ)と(Ⅱ)を取り入れた基準が多用されて来た。一方、(Ⅲ)の基準として、近年β-Divergence やγ-Divergence 等の統計的情報量が研究され、ロバスト推定などに応用されてきた。しかし、ビッグデータの普及により膨大なサブプロセスの条件・状態組合せを含む大規模対象系のモデリング需要に対して、これらは計算コストがかかり過ぎて実用にならない。データ分布を陽に扱わずとも、適切にモデルをマイニングする原理の確立が必要であり、本研究では局所モデル推定結果の統計的不偏性や一致性を満たしつつ、かつ実施項目(2)において高速探索を実現し易い性質の原理を模索した。実施項目(2)では、この結果を受けて高速探索原理を検討した。

実施項目(3)(4)では、実データの性質を踏まえた各種アルゴリズムを検討した。特に観測対象系における各サブシステム（データクラスタ）の粗密さにかかわらず、見落としが少なくかつ高速な探索が可能なアルゴリズム検討を行った。また、データの観測対象系に含まれるサブシステム（データクラスタ）に無関係な変数や外れ事例の多少も結果に影響を与えるので、それぞれに対応可能なアルゴリズムの検討と検証を行い、必要に応じて(1)(2)にフィードバックした。ここでは実例台として、特に心疾患を中心とする医療分野のデータを取り上げ、種々の原因疾患別に異なる発生分布をする心疾患について、それぞれ生起ダイナミクスのメカニズムを表すモデルのマイニングを行った。

４．研究成果

実施項目(1)については、従来の統計的・情報論的なモデル選択基準では、膨大なサブプロセスの条件・状態組合せを含む大規模対象系のモデリング需要に対して、計算コストがかかり過ぎて実用にならないことから、単純なデータのランダムサブサンプリングとアンサンブルモデリングを中心とする原理を探求した。ランダムサンプリングは対象データの大きさにかかわらず、定数時間で実行することが可能であり、かつ大量のランダムサンプリング行う場合でも並列計算処理化が容易で高速な処理が可能である。また、ランダムサブサンプリングによる統計的精度の低下をアンサンブルモデリングによって補うことを考えた。両原理の組合せにより、データ全体を対象としてモデル発掘を行う場合に比してモデル精度を落とすことなく、かつ遥かに高速にモデルマイニングを行うことができる。

実施項目(2)では、実施項目(1)で考案したランダムサブサンプリングを行う際の効率的でかつデータが持つ統計的及び情報論的特徴を損なうことのないサンプリングを可能にする手法やその条件を探求した。特にサンプリング結果から計算幾何学的原理によってデータが存在する各部分空間や各領域における統計的性質を高精度かつ効率的にモデル化可能な半空間データ質量という概念とそれを計算する手法を確立した。これによって、大規模で複雑な分布を有するデータから、非常に高速にデータ各部の特徴を表す統計的及び情報論的なモデルを発掘することが可能となった。また、この原理、手法を機械学習で標準的な大規模ベンチマーク問題を対象としたデータクラスタリング及びデータ異常値検知問題に適用し、大規模データに関して従来手法を遥かに超える計算速度と高いモデリング精度を達成可能であることを確認した。この研究は協力研究者であるオーストラリア連邦大学の研究チームと合同で行われ、その成果を機械学習分野の世界トップジャーナルである Machine Learning に主な発表論文〔雑誌論文〕①、同じくデータマイニング分野の世界主要国際会議の１つである PAKDD2015 に〔学会発表〕①として発表した。

さらに実施項目(2)では、実施項目(1)で考案したランダムサブサンプリングに基いて、データ事例間の類似性尺度を前述のデータ質量を基に計算する手法の開発を行った。これによって、大規模で複雑な分布を有するデータから、非常に高速にデータの部分部分各々においてその統計的及び情報論的な特徴を反映した類似性尺度を自動的に定義することが可能となった。また、この原理、手法を機械学習で標準的な大規模ベンチマーク問題を対象としたデータ分類問題に適用し、大規模データに関して従来手法を遥かに超える計算速度と高いモデリング精度を達成可能であることを確認した。この研究も協力研究者であるオーストラリア連邦大学の研究チームと合同で行われ、その成果をデータマイニング分野の世界トップ国際会議である ICDM2014 及び AIRS 2015 に主な発表論文〔学会発表〕①、②として発表した。

以上と平行して、実施項目(3)及び(4)については、主要な医療分野の１つであり日本人の死因の第２位である心疾患を適用分野として選び、国立循環器病研究センターの研究グループの協力を仰いで心疾患臨床データの解析に適用した。国立循環器病研究センターに過去入退院した患者の膨大な履歴データを対象として、実施項目(1)、(2)の原理及び手法を適用するアルゴリズムを開発、改良実装し、心疾患臨床データから原因疾患別に異なる発生分布をする心疾患それぞれの生起ダイナミクスを表すモデルのマイニングを行った。この結果、心疾患の生起メカニズムに関し、医学分野で従来知られていなかった新しモデルの発見に至っている。その成果の一部を主な発表論文〔雑誌論文〕②において発表しており、さらにより詳細な結果を本研究期間終了後になるが別途医学論文誌に投稿する手続きを進めている。

以上、全体を通じて当初の計画通りの成果を得ることができ、さらに実適用例題の医療分野においては、萌芽的研究段階であるにもかかわらず、予想以上に実践的な研究成果を得ることができた。

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 Bo Chen, Kai Ming Ting, Takashi Washio and Gholamreza Haffari, Half-space mass: a maximally robust and efficient data depth method, Machine Learning, Vol.100, No.2 (2015) pp.677-699

② Masafumi Kitakaze, Masanori Asakura1, Atsushi Nakano, Seiji Takashima and Takashi Washio, Data Mining as a Powerful Tool for Creating Novel Drugs in Cardiovascular Medicine: the Importance of a “Back-and-Forth Loop” between Clinical Data and Basic Research, Cardiovascular Drug and Therapy, Springer, Vol.29, No.3 (2015) pp.309-315

〔学会発表〕（計　３件）

① Sunil Aryal, Kai Ming Ting, Jonathan Wells and Takashi Washio, Improving iForest with relative mass, Proc. of PAKDD2014: The 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Advances in Knowledge Discovery and Data Mining, Lecture Notes

in Computer Science Vol.8444, 2014, pp.510-521
② Sunil Aryal, Kai Ming Ting, Gholamreza Haffari and Takashi Washio, mp-dissimilarity: A data dependent dissimilarity measure, Proc. of ICDM2014:IEEE International Conference on Data Mining, DM570, 2014
③ Sunil Aryal, Kai Ming Ting, Gholamreza Haffari and Takashi Washio, Beyond tf-idf and cosine distance in documents dissimilarity measure, Proc. of AIRS 2015: The 11th Asia Information Retrieval Societies Conference; In Information Retrieval Technology of the series Lecture Notes in Computer Science: Springer International Publishing, Vol.9460 (2015) pp 400-406

６．研究組織

(1)研究代表者

鷲尾　隆 (WASHIO, Takashi)
大阪大学・産業科学研究所・教授
研究者番号：００１９２８１５

(3) 連携研究者

清水　昌平 (SHIMIZU, Shohei)
大阪大学・産業科学研究所・准教授
研究者番号：１０５０９８７１

河原　吉伸 (KAWAHARA, Yoshinobu)
大阪大学・産業科学研究所・准教授
研究者番号：００５１４７９６