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　 ６ 月　 ３ 日現在

機関番号：１３７０１
研究種目：若手研究(B)
研究期間：2013～2015
課題番号：２５８７０３２２

研究課題名（和文）多様な補助知識を利用する高速な統計的機械学習アルゴリズム

研究課題名（英文）Statistical Machine Learning with Heterogeneous Auxiliary Information

研究代表者
志賀　元紀（SHIGA, Motoki）

岐阜大学・工学部・助教

研究者番号：２０４３７２６３

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：近年の産業・科学において様々な情報を同時計測できる状況が増えてきたため、計測・観測データ量が膨大化する傾向にある。そのため、こうした膨大な情報の背後に潜む簡潔な法則を発見する統計的機械学習法の需要が高まってきた。本研究では、外部データベースや熟練者による補助知識を組み合わせるデータ解析手法を新たに開発した。特に、ネットワーク構造情報やグループ情報を取り入れる手法の研究に従事した。また、得られた方法・知見に基づき、ゲノム科学・臨床研究・材料科学におけるデータ解析にも取り組んだ。

研究成果の概要（英文）：Recent developments of measuring engineering enable us to simultaneously monitor multiple variables and then the common size of datasets has been increasing. Thus finding essential simple rules hidden in such huge datasets becomes important. This research project has developed efficient clustering and matrix/tensor factorization methods by combining auxiliary information provided from databases and experts. Among a lot of data structures of auxiliary information, this project focused on auxiliary group and network structures. These results were also applied to data analysis on genome science, medical research, and material science.

研究分野：統計的機械学習

キーワード：行列分解　クラスタ解析　スパース正則化　変分ベイズ学習

１．研究開始当初の背景

近年の産業・科学において様々な情報を同時計測できる状況が増えてきたため、計測・観測データ量が膨大化する傾向にある。このようなデータ量の増加のみならず説明変数(特徴量)の数および多様性も増加しており、こうした複雑かつ膨大なデータを取り扱うためのデータマイニング・機械学習周辺の技術発展が渇望されるようになってきた。特に、スケーラビリティのあるデータ解析を実現するためには、調べたい現象に隠れた本質的な構造を抽出し、現象と密接に関係ある変数を選択することが基本的かつ重要な課題の一つとされる。例えば、様々な条件で細胞内の遺伝子発現量を網羅計測する場合、細胞としてのシステム全体の特性を決めるような遺伝子群を同定することが非常に重要な課題である。

ところで、新しい法則を発見したい多くの場合、新規に計測・観測・収集されたデータのみではなく、データベースや熟練者などによって提供される補助知識が有効な情報となりうる。これらを有効利用するために、主対象となる観測データセットと補助情報をいかに効率良く組み合わせるかが一つの重要な課題とされる。しかしながら、複数のデータを組み合わせて規則発見やモデルの学習を行う場合、単一のデータを扱う場合よりも複雑な最適化問題を解いて学習アルゴリズムを構築する必要がある。特に、大規模データを対象とする場合、高速な学習アルゴリズムが必要となるため近年も盛んに研究されている。

２．研究の目的

本課題では、大規模かつ雑音が多い観測データに潜在する規則（クラスタ構造や低ランク構造などの単純な構造）の発見を行う機械学習法の研究に従事する。特に、グループ構造やネットワーク構造を含む多様な補助情報を効率的に取り入れることによって雑音の影響を軽減させ、正確なデータ解析を実現する。前述した通り、異なる補助情報を組み合わせることによって、データからパラメータを最適化する問題が難しくなることが多い。そこで、劣勾配法などの手法を取り入れて、パラメータ最適化の高速化も念頭においた研究に取り組む。

また、研究期間内において、本課題の成果や得られた知見に基づく実データ解析への応用を念頭においている。より具体的には、網羅的な遺伝子発現量の計測データを用いた遺伝子機能の同定などのゲノム科学分野への応用や、走査電子顕微鏡データ解析などの材料科学分野への応用に焦点をおく。その際、従来法を単に流用するだけでなく、応用問題の個別の設定にあわせて手法の適切な拡張を目指す。

３．研究の方法

(1) 補助情報を用いる教師なし学習法

マイクロアレイや RNA-seq などの網羅的な遺伝子発現データを用いて遺伝子機能グループを同定する研究がこれまでも行われているが、観測標本数が少なく、かつ、観測雑音が大きいために誤った結論を導く可能性がある。こうした問題に対して、遺伝子ネットワークなどの既知の補助情報を用いる有効性が、申請者らが提案した統合的クラスタ解析法の研究成果などから示されてきた。本課題では、申請者らの従来研究をさらに発展させ、共クラスタ解析および行列分解において補助情報を導入する手法を構築した。補助情報を用いるために、確率モデルや正則化モデルの新規設計などに取り組み問題解決にアプローチした。

(2) 教師あり学習における特徴選択

疾患患者の生存時間解析をはじめとする臨床データ解析において、多数の特徴量（入力）と目的変数（出力）がペアとなった形で観測される。こうした観測において、どの特徴が目的変数に寄与するかを明確に理解できていない状況では、とりあえず網羅的にデータが記録される。こうした状況においても、教師あり学習法を用いれば、特徴量から目的変数の予測や関係式の推定を行えるものの、目的変数と関係ない特徴が悪影響を及ぼし、推定の精度が悪化する問題ある。本課題では、以下の 2 つの状況 ①条件付き確率密度関数のノンパラメトリック推定、②生存時間解析のモデル（Cox 比例ハザードモデル）における特徴選択問題にアプローチした。

(3) 実応用データ解析のための機械学習法

具体的に取り組んだ応用は、①ゲノム科学・臨床研究分野、②材料科学分野におけるデータ解析である。応用分野①において、遺伝子発現量データに関しては、米国 NCBI が一般公開しているデータベース Gene Expression Omnibus(GEO)を利用し、また、臨床研究データに関してはデータ解析コンペティション DREAM Challenge が提供するデータを用いた。一方、応用分野②の材料科学分野のデータに関しては、走査型電子顕微鏡の計測データを共同研究者から提供していただいた。これらの応用課題において、上述したクラスタ解析や行列分解や特徴選択などの研究に関わる解析法を中心として新規データ解析手法の構築を行った。

４．研究成果

(1) 補助情報を用いる教師なし学習法

① ネットワーク構造を補助情報として用いる共クラスタ解析法

がん細胞には多様な種類が存在し、各細胞の遺伝子の振る舞いも多様である。様々なが

ん細胞の遺伝子発現量データから遺伝子機能クラスタを同定する際には、同時にがん細胞クラスタも同定する必要があり、つまり、通常のクラスタ解析（単方向クラスタ解析）ではなく共クラスタ解析が望ましい。本研究では、少ない標本数のデータから生物学的に合理的な結果を導けるように、既知の遺伝子ネットワーク構造を取り入れる共クラスタ解析法を提案した。この手法は、遺伝子ネットワーク構造と発現量を組み合わせる確率モデルの変分ベイズ学習に基づくものであり、実データを用いた性能検証により有効性を示し、その成果を国際ワークショップ MultiClust2013 に発表した。

② グループ補助情報を取り入れる行列（テンソル）分解

文書データや遺伝子発現量などの多くの観測データが行列形式でまとまっており、これを時系列に観測するなど条件を加えればテンソル（多次元配列）のデータ構造にまとめられる。こうした観測データのサイズは膨大であっても、その要素が少ない因子・グループにまとめられるために、小さいサイズの行列のテンソル積によって観測データを近似できる。こうした潜在的な要素（低ランク行列）で元の観測データを近似する解析は行列分解（テンソル分解）とよばれる。本研究では、オーバーラップのあるグループ情報を外部補助情報として利用する方法を構築した。ここでいうグループとはノード（事例）の部分集合のことであり、また、グループのオーバーラップとは 1 つのノードが複数のグループに所属することである。グループ補助情報をモデル学習法に利用するため、仮定「同じグループに所属する観測値やパラメータは近い値になる」を満たすように、正則化項を導入した。さらに、このモデルのパラメータ最適化アルゴリズムを、階層的交互最小二乗法(HALS 法)に基づき導出した。人工データおよび実データを用いた数値実験によって、最適化の収束速度が従来法よりも速いこと、また、グループ補助情報の利用が有効であることを示すことができた。本研究成果を学術雑誌 IEEE TKDE などで発表した。現在、本提案法をより複雑なテンソル構造の分解に拡張すること考えており、近々、学術雑誌論文などの成果にまとめる予定である。

(2) 教師あり学習における特徴選択法

① スパース加法モデルによる特徴選択および条件付き確率密度関数の推定

条件付き確率密度関数のノンパラメトリック推定と同時に特徴選択を行う手法を構築した。提案法は基底関数モデルに基づく手法であり、基底関数の重み係数推定と同時に特徴選択するために、重み係数にスパース正則化項を導入するモデルを新たに設計し、その最適化アルゴリズムも導出した。人工データやロボット制御シミュレーションデータなどによるベンチマーク実験によって性能検証した結果をまとめて、国際会議 ECML/PKDD2015 および学術雑誌 Machine Learning にて発表を行った。

② 生存時間解析における特徴選択法

疾患が発生してからの生存時間を予測する生存時間解析において、臨床データと生存時間の関係をモデル化するには、Cox 比例ハザードモデルが用いられる。L1 ノルムに基づくスパース正則化によって、予測に重要な臨床項目を絞り込む解析が従来から提案されているが、本来必要のない特徴を余分に選ぶ傾向にある。こうしたことから、スパース正則化項付き学習で絞り込んだ特徴群に対して、前向き特徴選択を実行することで、余分な特徴を取り除く方法を試みた。データ解析コンペティション DREAM 9.5 Prostate Cancer DREAM Challenge にて提供された約 1000 人の前立腺癌患者の臨床データからモデル学習を行い、約 500 人の患者の生存時間を予測するタスクを行ったところ、正確な予測を実現でき、コンペティションで入賞した。また、提案方法をオンライン論文誌 F1000 Research にて発表する予定である。

(3) 実応用データ解析のための機械学習法

① 電子顕微鏡データを用いた電子状態の自動マッピング法

走査型電子顕微鏡の分光スペクトル計則技術が発達しており、評価試料平面の各点での電子状態を反映するエネルギー損失スペクトルを高速に計測できるようになってきた。このデータの大雑把なサイズは、例えばスペクトル検出チャネル数が 2000、平面上の地点数が 100×100 であり、合計 2000 万点と非常に膨大であるため、電子状態マッピングの自動化が必須である。これに対して、従来から主成分分析(PCA)を拡張した手法が使われてきたが、物理的に不自然なモデルであるため、不自然な結果を導く問題が指摘されてきた。これに対して、本研究では非負値行列分解に基づく手法によりアプローチし、特に直交制約に基づき不自然な重なりを解消する方法、ARD 事前分布に基づき試料中に含まれる電子状態の種類数を自動決定できる方法を構築した。これらの研究成果を学会発表および学会誌にて発表した。（現在、学術雑誌論文に投稿中である。）

②グループ情報を用いた遺伝子発現量解析

以前から取り組んでいた研究、細胞状態・実験条件によって発現量レベルが異なる遺伝子を高速に検知する手法に取り組んだ。この手法をさらに発展するために、遺伝子発現量の有意差検知を、1 つの遺伝子ではなく遺伝子グループ単位で行う関連手法をサーベイし、学術雑誌 IEEE/ACM TCBB にて発表を行った。

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 4 件）

① Motoki Shiga, Voot Tangkaratt, Masashi Sugiyama, Direct Conditional Probability Density Estimation with Sparse Feature Selection, Machine Learning, vol.100, no.2, pp.161-182, 2015, 査読有り,
doi: 10.1007/s10994-014-5472-x

② Motoki Shiga, Hiroshi Mamitsuka, Non-negative Matrix Factorization with Auxiliary Information on Overlapping Groups, IEEE Transactions on Knowledge and Data Engineering, vol.27, no.6, pp.1615-1628, 2015, 査読有り,
doi: 10.1109/TKDE.2014.2373361

③ Mitsunori Kayano, Motoki Shiga, Hiroshi Mamitsuka, Detecting Differentially Coexpressed Genes from Labeled Expression Data: A Brief Review, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(1), pp.154-167, 2014, 査読有り
doi: 10.1109/TCBB.2013.2297921

〔学会発表〕（計 10 件）

① Motoki Shiga, Shunsuke Muto, Kazuyoshi Tatsumi, Koji Tsuda, Nonnegative Matrix Factorization for Spectral Imaging Data Analysis, International Meeting on “High-Dimensional Data Driven Science” (HD3-2015), Mielparque Kyoto, Japan, Dec. 14-17, 2015.

② Motoki Shiga, Voot Tangkaratt, Masashi Sugiyama, Direct Conditional Probability Density Estimation with Sparse Feature Selection, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD2015), Porto, Portugal, Sept 7-11, 2015.

③ 志賀元紀，多様な特徴量に基づく目的変数予測のための統計的機械学習，日本化学会情報科学部会主催 第二回若手の会，日本化学会化学会館，東京，11 月 29 日，2014.（招待講演）

④ 志賀元紀，機械学習に関する最適化問題，日本オペレーションズ・リサーチ学会・中部支部研究会，名古屋工業大学，10 月 25 日，2014.（招待講演）

⑤ 志賀元紀，津田宏治，武藤俊介，STEM-EELS スペクトラムイメージ解析のための非負値行列分解法，日本金属学会 2014 秋期講演大会，名古屋大学東山キャンパス，9 月 24 日-26 日，2014.

⑥ 志賀元紀，杉山将，特徴選択を同時にできる条件付き確率推定法，ニューロコンピューティング研究会，岐阜大学サテライトキャンパス，岐阜，12 月 21 日，2013.（電子情報通信学会，信学技報，vol. 113, no. 374, NC2013-56, pp. 17-22, 2013 年 12 月.）

⑦ 志賀元紀，杉山将，スパース加法モデルに基づく条件付き確率推定法，第 16 回情報論的学習理論ワークショップ（IBIS2013），テクニカルトラック，ポスター，T-08，東京工業大学 蔵前会館，東京，11 月 11 日-13 日，2013.（電子情報通信学会，信学技報，vol. 113, no. 286, IBISML2013-43, pp. 53-60, 2013 年 11 月.）

⑧ Motoki Shiga, Hiroshi Mamitsuka Variational Bayes Co-clustering with Auxiliary Information, Proceedings on the 4th MultiClust Workshop on Multiple Clusterings, Multi-view Data, and Multi-source Knowledge-driven Clustering (MultiClust2013), pp.1-4, Chicago, Illinois, USA, August 11-14, 2013.

〔その他〕
学会誌解説記事

① 武藤 俊介，志賀 元紀，巽 一厳，津田宏治，ナノ電子顕微分光における情報処理技法の応用，日本セラミックス協会「セラミックス」，50(7), pp.527-530, 2015.（査読なし）

６．研究組織
(1)研究代表者
志賀　元紀（SHIGA, Motoki）
岐阜大学・工学部・助教
研究者番号：20437263