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　２４　日現在

機関番号：　３２７０６
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５３３００４５

研究課題名（和文）統計的学習理論と凸最適化アルゴリズムに基づく大規模データの自動分類法に関する研究

研究課題名（英文）Big Data Classification Methods and Applications Based on Statistical Machine Learning and Convex Optimization

研究代表者

小林　学（KOBAYASHI, Manabu）

湘南工科大学・工学部・教授

研究者番号：８０３０８２０４

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究を通して，大規模データに対する統計的手法並びに凸最適化に基づく自動分類法を応用することにより，広範な諸問題に対して精度の高い解を効果的に求めることが可能であることを示した．
具体的にはデータに秘匿性を持たせて学習を行うプライバシー保護分散処理問題，ECサイトにおける潜在クラスの解析問題，動的再構成回路の設計問題，L1最適化による文書分類問題，CARTを用いた無ひずみデータ圧縮，授業への出欠情報を用いた隠れ属性モデル解析問題，ランダムマルコフフィールドを用いた大規模故障診断問題，大規模プログラミング編集履歴取得・可視化，等々に対するアルゴリズム並びに解析手法を提案し，それらの有効性を示した．

研究成果の概要（英文）：Applying classification methods based on the statistical machine learning and convex optimization for big data, we showed that it was possible to obtain efficiently the high precision solutions for wide range of various problems.
Specifically, we proposed algorithms and analysis methods, and showed the effectiveness for the following problems:
(1)privacy preserving distributed calculation problem for the case which some parties have different secret data, (2)latent class model analysis problems of EC site or institutional research, (3)dynamic reconfiguration circuit design problem, (4)document classification problem based on L1 optimization, (5)lossless data compression using CART, (6)fault-diagnosis problem using markov random field, and (7)programming edit history acquisition and visualization problem for many students.

研究分野：　学習理論

キーワード：　学習理論　凸最適化　統計的モデル　ビッグデータ解析　隠れ属性モデル　メトリックラーニング　I-S cover

１．研究開始当初の背景

ブログや SNS などコミュニケーションを目的とする WEB サービスや，クラウドを用いたファイル同期サービスなど，日常的に大規模なデータをクラウド上で処理するサービスが急速に発展している．これらビッグデータはデータセンタ内の多数のサーバ上に配置され，必要に応じて入力・更新・出力が行われている．これらのデータは顧客やユーザの知識や行動，興味などに関する多くの情報を含んでおり，その解析により特徴を自動抽出することによって顧客やユーザへのフィードバックに活用することが可能となる．一方従来から統計学や学習理論により，このようなデータの解析法は種々研究が行われている[①]．与えられた学習データを用いて，新規データがどのカテゴリに所属するかを自動的に判別する自動分類法は，回帰問題と並んで非常に重要な役割を演じている．スパムメールフィルタなどは，自動分類法の直接的な応用である．

また昨今マルチコア及びマルチコンピュータシステムによる並列化処理を導入することにより，高度な処理を高速に処理することが可能となってきている点も重要である．

２．研究の目的

近年クラウドコンピューティングの急速な普及により，クラウド上では日常的に大規模なデータが処理されるようになった．このようなビッグデータは目的とするサービスのための処理以外に，データそのものを解析することにより特徴抽出や分類，回帰など様々な応用へ利用することが可能となってきている．

本研究では各サーバ間での並列処理と通信を繰り返すことにより，各サーバが個別に保有しているデータを高速かつ統一的に解析する手法の開発を目的とする．特にビッグデータに対する自動分類法とこれを用いた応用に焦点を当て，統計的手法と凸最適化の手法の両者を組み合わせることにより，広範な諸問題に対して精度が高くかつ高速なアルゴリズムを導出し，有効性の評価を行う．

３．研究の方法

本研究では初年度に調査した自動分類法における様々な統計的手法並びに凸最適化手法を元に，その両者の長所を用いることによって広範な諸問題に対して精度が高くかつ効率的なアルゴリズムの提案並びに解析手法の導出と評価を行った．具体的に初年度の調査結果を元に取り組んだ諸問題は以下である．

(1) プライバシー保護分散処理問題
(2) EC サイトにおける潜在クラスの解析問題
(3) 授業への出欠情報を用いた潜在クラスモデル解析問題
(4) L1 最適化を用いたメトリックラーニングと文書分類問題への応用
(5) 動的再構成回路の設計問題
(6) CART を用いた無ひずみデータ圧縮
(7) ランダムマルコフフィールドを用いた大規模故障診断問題
(8) 大規模プログラミング編集履歴取得・可視化

それぞれの研究において文献調査の他，数式による理論的な解析とアルゴリズムの導出，さらには統計的生成モデルによる数値計算実験並びに実データを用いた大規模計算機実験による評価と考察を実施した．

次節においてそれぞれの分野における成果を示す．

４．研究成果

本節では取り組んだ各問題の内容と成果を解説する．

(1)プライバシー保護分散処理問題[雑誌論文①]

本研究ではプライバシー保護を目的とした分散処理学習問題を扱っている．これは複数のサーバがそれぞれ異なるデータを保持したもとで，それぞれの保持するデータを互いに公開，共有することなく，協力して全てのデータを用いた場合と同等の分析を行うというものである．本研究では分散処理によって最小二乗推定量を学習する新たな分散計算プロトコルを提案し，プライバシーについてプロトコルの安全性を評価している．このとき推定量の最適性並びに収束性は保証されるが，収束までの時間は数値実験により評価を行った．なお本研究の結果はリッジ回帰や Lasso への応用も可能と考えており，今後も継続して取り組む予定である．

(2)EC サイトにおける潜在クラスの解析問題[雑誌論文②，学会発表①～③]

本研究では EC サイトの大量の購買履歴及び商品の閲覧履歴に対して統計的モデルを設定し，商品及び顧客の見えない属性（潜在クラス）を分類する手法の確立を行った．具体的には顧客及び商品には見えない潜在クラスが存在するものと仮定し，これらはそれぞれ多項分布で発生するものとする．一方これらの商品及び顧客の潜在クラスの組み合わせで，購買あるいは商品の閲覧行動を確率的に行う統計的モデルの構築を行った．さらに，これらを学習する手法として，ＥＭアルゴリズム及びその改良として変分ベイズ法の適用を行い，高速かつ高精度に尤度あるいは事後確率の高いパラメータを求める手法の提案を行った．また結果的に得られた潜在クラスに対して，顧客及び商品のどの属性が主に影響するかを自動分類の手法を用いて解析するアルゴリズムの提案を行い，有効性の評価を行った．本手法は他のマーケティング分野への応用を考えることができ，さらに

幅広い応用を持つと考える．

(3)授業への出欠情報を用いた潜在クラスモデル解析問題[学会発表④]

本研究では授業における学生の出欠情報を用いて，学生のクラスタリング及び教員の特徴づけを行う手法を提案した．具体的には学生及び授業には見えない属性（潜在クラス）が存在することを仮定し，この潜在クラスの属性値の組み合わせによって授業の出席率が確率的に決定する確率モデルを提案した．さらにこの確率モデルのパラメータを機械学習により推定する方法を示し，評価を行った．結果的に，「学生の学習支援や教育指導に対する改善」並びに「各教員の教授活動や学生指導に対する改善」に利用可能かどうか考察を行った．本手法は(2)の直接的な応用として位置づけることができ，機械学習の手法を教育への IR 活動に有効活用できることを明らかにした重要な成果である．今後雑誌論文に投稿予定である．

(4)L1 最適化を用いたメトリックラーニングと文書分類問題への応用[雑誌論文③，学会発表⑤～⑦]

データの統計的特徴を考慮した距離構造を学習する方法としてメトリックラーニングが知られており，そのための様々な手法が提案されている．メトリックラーニングはマハラノビス距離におけるマハラノビス行列（以下，計量行列）を学習するための手法であるが，パラメータ数が入力データの次元数の 2 乗に比例することが知られている．加えて，学習に要するデータの数も同様に増加してしまうため，高次元データを用いた場合には非常に多くのデータを用意する必要がある．本研究では，計量行列のパラメータ数を減少させるための方法として L1 正則化に基づくアプローチを採用し，凸最適化手法である ADMM(Alternating Direction Method of Multiplier)を用いた最適な計量行列の導出方法を示した．また正則化の効果について解析を行ない，固有値との関係を明らかにしている．さらに提案手法を高次元，スパースなデータセット，ならびに低次元，密なデータセットとして文書分類問題に適用し，その有効性を示した．

本研究は計量行列に対する適切な評価関数を系統だって記述することにより，ADMM を用いて効率的に所望の解を導出できることを明らかにしており，重要な研究である．また得られた計量行列の意味に着目し，文書に対する単語の役割の解析に活用できると考えており，今後のさらなる発展が期待できる．

(5)動的再構成回路の設計問題[雑誌論文④～⑨，学会発表⑧，⑨]

本研究では最適化手法を用いた動的再構成可能な回路の設計を扱っている．FPGA のようにプログラムによって動的に再構成可能な回路は益々重要となっているが，近年 Double Gate Carbon Nanotube Field Effect Transistor (DG-CNTFET)が回路素子として動的再構成回路の要素に成り得ることが発見された．さらに DG-CNTFET を用い，この素子へのコントロール信号を制御することによって，任意の回路を再構成可能な万能回路を設計することが可能であることが示された．本研究では DG-CNTFET の配置及びコントロール信号に対して，最適化手法を適用することにより従来手法よりも少ない回路規模及び消費電力で万能回路を構築可能であることを示した．さらに t 項までの多項式を実現する回路を構築し，少ない回路規模で多数の効果的な動的回路を再構成可能であることを示した．また実際に加算や減算等を行う ALU の設計まで行い，十分な素子数の減少度の解析を行った．

本研究成果は，従来用いられていない DG-CNT の絶大な効果を明らかにしており，価値の高い研究と考えている．今後はより複雑な回路設計に対して最適化を実施し，複雑な回路ほど本研究の効果が十分に高いことを明らかにしたい．

(6)CART を用いた無ひずみデータ圧縮[学会発表⑩]

本研究では単一あるいは複数の 2 分木を用いて，2 次元データである画像の無歪みデータ圧縮を行う手法を提案した．このとき画像圧縮の特性に着目し，すでに圧縮した箇所を説明変数とし，次に圧縮する色要素を目的変数として扱っている．具体的にはまず圧縮対象の画像について CART 及びランダムフォレストで学習を行い，学習結果である 2 分木を符号化する．次にこれらの 2 分木を用いて圧縮対象の各画素値の推定を行い，推定値と画素値との差分のみを圧縮する．復号側はすでに受け取った 2 分木及び画素を用いて次の画素値を推定し，その推定値と圧縮されている差分値を加えることにより元の画素値を復元する．結果的に，いくつかの画像に適用したときに PNG 形式よりも圧縮可能であることを示し，その有効性の評価を行った．従来は AR モデルの予測を用いていた箇所に，画素の重要度を用いた 2 分木による予測を用いている点で本研究は新規性が高い．また圧縮率の高さという点で，大変インパクトのある研究であると考える．GPU を用いた高度な並列化が容易である点から，実用的にも興味深い結果と考える．

(7)ランダムマルコフフィールドを用いた大規模故障診断問題[学会発表⑪]

本研究では，ネットワークで接続された分散システムに対して，それぞれのノード（PC でも良いし，CPU あるいはコアでも良い）がお互いに故障の有無についてテストを行う故障診断モデルを考える．このとき，テストの回数を少なくしつつ，故障の有無を誤る確

率を非常に小さくする必要がある．マルチプロセッサシステムあるいはマルチコンピュータシステムにおける故障ノードを発見する問題は，古くから様々なモデルに対して研究が行われてきた．F.P.Preparata らは，システムの要素である各ノードをそれぞれ別のいくつかのノードが独立に検査を行い，それらの結果からシステム中の全ての故障ノードを発見する故障診断モデル（以下 PMC モデルと参照する）を提案した．これは故障でないノード（正常ノードと呼ぶ）が他のノードの検査を行った結果は正しく，故障ノードが他のノードの検査を行った結果は信頼できないという故障診断モデルとなっている．

本研究では Preparata の故障診断問題を確率的故障診断問題として定式化し，さらにある有向グラフの構造を仮定するとこれがランダムマルコフフィールドとして評価可能であることを示した．さらにこの構造に対して最適な確率的診断法を示し，機械学習の手法を用いることによりロバスト性の解析を行ない，その有効性を示した．故障診断問題は古くからグラフ構造の問題として捉えられてきたが，確率的構造に着目し，統計的手法によって事後確率最大化基準及びこれの解析手法を用いることを明らかにしている点で，価値ある研究である．本研究の内容は故障モデルの拡張を行うことができ，より広いグラフ構造に対して適用可能であると考えている．

(8)大規模プログラミング編集履歴取得・可視化[学会発表⑫]

コンピュータシステムの高度化及びスマートフォンの急速な普及に伴い，ソフトウェア開発技術者の確保並びに質向上は大きな問題となっており，プログラミング教育の必要性は益々高まっている．一方クラウド上のビッグデータに対して機械学習やパターン認識技術などの研究が盛んに行われており，高度な解析が可能となってきている．本研究では初学者のプログラミング教育を対象に，WEB 上でプログラムの作成及び実行を可能とし，同時に詳細な学習者の編集履歴を取得するシステムを構築した．さらに学習者の編集履歴データベースに対してリアルタイムに可視化を行うことにより，学習者のプログラム作成状況やエラーの状況，あるいは編集履歴や実行結果等をリアルタイムに把握及び表示することが可能となった．さらにこのシステムを用いると可視化の効果により，学習者全員の実行結果一覧を抽出することによってクラス全体の理解度をリアルタイムに把握することが可能となった．

本研究は集めたプログラミングの編集履歴を可視化までしか行うことができていないが，今後はさらに自動分類手法を応用することにより，学習者の問題点や自動採点などを行うことを考えている．これにより，教師の負担を減らすのみならず，個々の学習者への素早い対応を補助するシステムへと発展させたい．

以上のように，自動分類法並びにその応用について，統計的手法及び凸最適化を系統立てて用いることにより，広範な分野に役立てることが可能であることを明らかにした．この点で本研究の目的を十分に達成したものと評価している．なお陽には記述していないが，本研究で扱っているアルゴリズムはほとんどの場合において並列化が可能であり，分散処理により高速化が可能であることを付記しておく．

統計的手法と凸最適化手法の両者は通常別々に扱われることが多く，評価尺度も異なっていることが通常である．ただし本研究を通して互いの長所を理解し，適切に両者を用いる重要性は改めて強調したい．それぞれの分野における今後の課題は各項目で挙げたとおりであるが，大きな研究課題として，ある程度モデルを限定した下で，確率的手法と様々な最適化手法の両者の同一性を明らかにすることが挙げられる．

＜引用文献＞

① C. M. Bishop, *Pattern Recognition And Machine Learning*, Springer, 2006.

５．主な発表論文等

〔雑誌論文〕（計１１件）

① 須子統太，堀井俊佑，小林 学，後藤正幸，松嶋敏泰，平澤茂一，″プライバシー保護機能を持つ線形回帰モデルにおける最小二乗推定量の分散計算法について，″日本経営工学会論文誌，Vol.65, No.2, pp.78-88, 2014. 査読有

② M.Goto, K.Mikawa, S.Hirasawa, M.Kobayashi and S.Horii, ″A New Latent Class Model for Analysis of Purchasing and Browsing Histories on an EC Site,″ Journal of Industrial Engineering & Management Systems, Vol.14, No.4, pp.1-12, Dec. 2015. 査読有

③ 三川健太，小林 学，後藤正幸，″教師あり学習に基づく L1 正則化を用いた軽量行列の学習法に関する一考察，″日本経営工学会論文誌，Vol.66, pp.230-239, No.3,2015, 査読有

④ M. Kobayashi, H. Ninomiya, Y. Miura and S. Watanabe,″Reconfigurable Dynamic Logic Circuit Generating t Term Boolean Functions Based on Double-Gate CNTFETs,″IEICE Trans.on Fundamentals,Vol.E97-A,No.5,pp.1051-1058,2014.査読有

⑤ H. Ninomiya, M. Kobayashi, Y. Miura and S. Watanabe,"Reconfigurable Circuit Design based Arithmetic Logic Unit using Double Gate CNTFETs," IEICE Trans.on Fundamentals, Vol. E97-A, No.2, pp.675-678, 2014.査読有

⑥ J. Kato, S. Watanabe, H. Ninomiya , M. Kobayashi and Y. Miura, "Circuit Design of 2-Input Reconfigurable Dynamic Logic Based on Double Gate MOSFETs with Whole Set of 16 Functions," Contemporary Engineering Sciences, Vol.7, no.2, pp.87-102, Feb.2014. 査読有

⑦ J.Kato, S.Watanabe, H.Ninomiya, M.Kobayashi and Y.Miura,"Circuit Design of Reconfigurable Dynamic Logic Based on Double Gate CNTFETs Focusing on Number of States of Back Gate Voltages," Contemporary Engineering Sciences,Vol.7, no.1,39-52, Jan.2014. 査読有

⑧ M. Kobayashi, H. Ninomiya and S. Watanabe, "Circuit Design of Reconfigurable Logic Based on Double-Gate CNTFETs," IEICE Trans. on Fundamentals, vol. E96-A, no.7, pp.1642-1644, Jul. 2013. 査読有

⑨ H. Ninomiya, M. Kobayashi and S. Watanabe, "Reduced Reconfigurable Logic Circuit Design based on Double Gate CNTFETs using Ambipolar Binary Decision Diagram," IEICE Trans. on Fundamentals, vol.E96-A, no.1, pp. 356-359, Jan. 2013. 査読有

⑩ K.Umezawa, T.Ishida, M.Aramoto, M.Kobayashi, M.Nakazawa and S.Hirasawa , "A Method based on Self-study Log Information for Improving Effectiveness of Classroom Component in Flipped Classroom Approach," International Journal of Software Innovation (IJSI), Vol.4, Issue 2, pp.17-32, March 2016. 査読有

⑪ 石田 崇, 小林 学,"大学における情報数理系科目のインタラクティブ教材の試作とその評価,"CIEC コンピュータ&エデュケーション, vol. 35, pp.75-80, 2013. 査読有

〔学会発表〕(計１３件)

① M.Kobayashi, M.Goto, T.Matsushima and S.Hirasawa,"Latent Class Model Analysis Based on the Variational Bayes," Proc. of 2016 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, Honolulu, Hawaii, USA, pp.125-128, March 2016.

② M.Goto, K.Minetoma, K.Mikawa, M.Kobayashi and S.Hirasawa, "A Modified Aspect Model for Simulation Analysis," Proc.of 2014 IEEE International Conferenceon Systems, Man, and Cybernetics, pp.1325-1330, SanDiego, California, Oct.2014. 査読有

③ M.Goto, K.Mikawa, M.Kobayashi, S.Horii, T.Suko and S.Hirasawa, "An Analysis of Purchasing and Browsing Historiesonan EC Site Based on a New Latent Class Model," Proc.of the 1st East Asia Workshop on Industrial Engineering, Hiroshima, Japan,2014. 査読有

④ 小林 学, 後藤正幸, 平澤茂一, "出欠情報による潜在クラスモデル解析," 経営情報学会春季全国大会予稿集, pp.185-188, May 2015. 査読無

⑤ K.Mikawa, M.Kobayashi, M.Goto, S.Hirasawa, "A Study of Distance Metric Learning by Considering the Distances between Category Centroids," Proc. of 2015 IEEE International Conference on Systems, Man, and Cybernetics,pp.1645-1650, Hong Kong, Oct.2015. 査読有

⑥ K.Mikawa, M.Kobayashi, M.Goto and S.Hirasawa,"A Proposal of L1 Regularized Distance Metric Learning for High Dimensional Sparse Vector Space," Proc. of 2014 IEEE International Conferenceon Systems, Man, and Cybernetics, pp.2000-2005, SanDiego, California, Oct.2014. 査読有

⑦ 三川健太, 石田 崇, 小林 学, 後藤正幸, 平澤茂一,"高次元かつスパースなベクトル空間における 11 正則化に基づく計量距離学習に関する一考察,"情報理論とその応用シンポジウム予稿集, CD-ROM, pp.525-529, Nov. 2013.

⑧ M. Kobayashi, H. Ninomiya, Y. Miura and S. Watanabe,″DRDLC Generating t-Term Boolean Functions Based on DG-CNTFETs,″ Proc.of 2014 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing,
Honolulu,Hawaii,USA,pp.81-84,March 2014. 査読有

⑨ Y, Miura, H. Ninomiya, M. Kobayashi and S. Watanabe, ″An Universal Logic-Circuit with Flip Flop Circuit Based on DG-CNTFET,″ Proc.of 2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp.148-152, August 2013. 査読有

⑩ 小林 学, 松島敏泰, 平澤茂一, ″回帰木を用いた画像の無歪みデータ圧縮,″ 情報理論とその応用シンポジウム予稿集, CD-ROM, pp.364-369, Nov. 2013. 査読無

⑪ M.Kobayashi, M.Goto, T.Matsushima and S.Hirasawa, ″Robustness of Syndrome Analysis Methodin Highly Structured Fault-Diagnosis Systems,″ Proc. of 2014 IEEE International Conference on Systems, Man, and Cybernetics,pp.2815-2821, SanDiego, California, Oct.2014. 査読有

⑫ 小林 学, 後藤正幸, 荒本道隆, 平澤茂一, “プログラミング編集履歴可視化システムとその実践,” 日本経営工学会2015年秋季大会予稿集, Nov. 2015. 査読無

⑬ K.Umezawa, M.Aramoto, M.Kobayashi, T.Ishida, M.Nakazawa and S.Hirasawa,″An Effective Flipped Classroom based on the Log Information of the Self-study,″ Proc. of the 3rd International Conference on Applied Computing & Information Technology (ACIT 2015), pp.263-268, Sep.2015. 査読有

〔図書〕(計１件)
① 後藤正幸, 小林学, 入門 パターン認識と機械学習, コロナ社, pp.72-168, 2014年4月

６．研究組織
(1)研究代表者
小林 学 (KOBAYASHI, Manabu)
湘南工科大学工学部情報工学科・教授
研究者番号：８０３０８２０４

(2)研究分担者
平澤 茂一 (HIRASAWA, Shigeichi)
早稲田大学理工学術院・名誉教授
研究者番号：３０１４７９４６

(3)研究協力者
吉本 昌史 (YOSHIMOTO, Masashi)