様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　８　日現在

機関番号：　６２６０３
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2016
課題番号：　２６２８００１０

研究課題名（和文）エッジヘビィデータ環境下におけるストリーミング計算用非線形フィルタ手法の研究

研究課題名（英文）Study on Nonlinear Filtering Method for Streaming Computing under Edge Heavy Data Environment

研究代表者

樋口　知之（HIGUCHI, Tomoyuki）

統計数理研究所・その他部局等・所長

研究者番号：７０２０２２７３

交付決定額（研究期間全体）：（直接経費）　　12,100,000円

研究成果の概要（和文）：ビッグデータの量の著しい増大スピードの一因である発生速度の増大は、センサー技術の向上と廉価化がもたらした産物である。情報システムの実世界との接点となる現場でのデータ発生頻度は増大の一途である。この現象はエッジヘビィデータ問題とも呼ばれている。ビッグデータはそのままクラウドへ輸送することは現実的でなく、その場で目的に応じたオンライン計算が必須である。本研究では、異なる特性をもつフィルタリング機能の優れた点を融合した手法の開発を目標とする。その結果、ストリーム計算の適用可能性を拡大する。

研究成果の概要（英文）：Increasing the rate of occurrence, which is one factor of the significant increase in the amount of big data, is a product brought about by improved sensor technology and lower cost. The frequency of occurrence of data at the site that is a contact point with the real world of information systems is ever increasing. This phenomenon is also called an edge heavy data problem. It is not realistic to transport big data as it is to the cloud, and online computation according to purpose is essential on the spot. In this research, we aim to develop a method combining the excellent points of nonlinear filtering with different characteristics. As a result, we extend the applicability of stream computing technology in machine learning technique.

研究分野：　情報学

キーワード：　時系列解析　逐次データ同化

１．研究開始当初の背景

（１）①ビッグデータの量の著しい増大スピードの一因である発生速度の増大は、センサー技術の向上と廉価化がもたらした産物である。例えば、ビデオ・サーベイランス（監視カメラ）を用いた犯人の特定や追跡は今後益々ますます効果をあげることが期待されているが、ビデオの時間・空間解像度の改善は、結果として大容量・高頻度データの産出につながっている。セキュリティばかりでなく工場の生産ラインなどでも、ビデオや画像を利用した管理・検査が大規模に導入されており、情報システムの実世界との接点となる現場でのデータ発生頻度と容量は増大の一途である。この現象はエッジヘビィデータ（Edge-Heavy Data）問題とも呼ばれている。

②通常、エッジに蓄積されるデータすべてが有益な情報ではなく、むしろそのほとんどがゴミ（“ Exhaust Data ”（廃棄データ）とも呼ばれる）であり、そのまま輸送（転送）していてはコストが相当増える。そもそも莫大な通信量が通信システム全体に与える負荷からして問題である。この ICT インフラのリソースを消耗する意味でのビッグデータの消耗性が、ビッグデータ関係者をして「ビッグデータは価値密度が低い」と言わしめるのである。データ利用の主目的に合った価値がデータ内に間欠的に散在するため、ビッグデータの場合、価値の総和はかなり増えるが、総データ量で総価値を割り算した「価値密度」は極めて小さくならざるをえない。このため、ストリーム計算（Stream Computing）のような、エッジですぐデータを加工する計算技術が今後は重要になる。

（２）①ストリーム計算の研究は、現在、機械学習分野で中心的に行なわれている。その研究動向は以下の 2 点で特徴づけられる。一つは、問題の多くが判別問題を取り扱っていることである。そこでの学習器は、時間に依存した巨大な次元の説明変数ベクトル（入力ベクトル）と、一般には時間に弱く依存すると仮定するパラメータ（重み係数）ベクトルおよび多くはバイナリーである目的変数（出力）で構成される。もう一つの特徴は、時変パラメータの推定が勾配法および L1 正則化をベースとするような最適化手法により点推定で実現されていることである。

②ストリーム計算はその機能から、ビッグデータ環境下でのオンライン・フィルタリングと位置づけられる。巨大次元の時系列データにかかわるオンライン・フィルタリングは、機械学習分野より先んずること約 20 年、気象・海況予報の領域で継続的に研究開発がなされてきた。フィルタリングの手法は逐次データ同化と呼ばれ、アンサンブル・カルマンフィルタがその先駆的研究として名高い。アンサンブル・カルマンフィルタ同様の、アンサンブル近似に基づく非線形フィルタとしては、すでに粒子フィルタが提案されていたが、粒子フィルタは一般に非常に多くのアンサンブルメンバ数を必要とするため、逐次データ同化の領域ではあまり利用がすすまなかった。両者の折衷方策として融合粒子フィルタが我々の研究チームによって提案され、簡便かつ実用性の高さから応用例が増えてきている。近年は、より計算量を減ずる方向に研究がすすみ、アンサンブル変換カルマンフィルタのように、小数のアンサンブルメンバ数の次元スケールの分散共分散行列の計算でもって近似簡略化する実用的な研究が盛んである。

（３）ストリーム計算の要素技術の研究は、機械学習領域と気象予報といった別領域で独立して進んでいるのが現況である。機能的には共にオンライン・フィルタリングであるため、両者の優れた点を生かした新しい手法の提案が可能である。ヒントは、ストリーム計算の学習器で利用する最適化関数の解釈にある。最適化関数は３つの項の和から成り立ち、一つは観測モデル、一つはパラメータベクトルの時間に関する平滑（smoothness）事前情報の項と見なせる。残りの項である、パラメータベクトル単独への拘束条件は、擬観測データ（時間一定の値）との違いを表す観測モデルと解釈可能であり、これらのことから学習器は、パラメータベクトルを状態ベクトルとする非線形非ガウス型の状態空間モデルとして表現可能である。

２．研究の目的

（１）まず、目的の全容を述べる。ビッグデータは、情報システムの観点からは、エッジと言える計測・観測の現場で大量生産されており、そのままクラウドへ輸送することは現実的でなく、その場で目的に応じたオンライン計算が必須である。その目的のために、機械学習分野ではストリーム計算と呼ばれる研究分野が大きな注目を浴びている一方、大規模な次元の観測ベクトルデータの処理に関しては気象・海況予報分野において逐次データ同化手法の研究がこの 20 年間継続的になされてきた。本研究では、この異なる特性をもつフィルタリング機能の両者の優れた点を利用した手法の開発を目標とする。その結果、ストリーム計算の適用可能性を拡大するとともに、新しい逐次データ同化法にもとづく簡便なシミュレータ（エミュレータと呼ばれる）技術を生み出す。

（２）①機械学習領域のストリーム計算を状態空間モデルの枠組みで表現し、最適化問題を統計的推測問題に拡張する。それにより、次の二つの方向性でもって研究を進める。

②一つは、逐次データ同化のエッセンスを機械学習分野のストリーム計算に導入することにより、新しいストリーム計算の枠組みを展開する。具体的には、ストリーム計算ではアドホックに定めていたメタパラメータを、データ適用的に決定するアルゴリズムを提案する。あるいは、逐次データ同化法を適用し、そのアンサンブルから最適化による推定

とは異なる特性をもつパラメータベクトルを算出する方策を検討する。
③もう一つの流れは、ストリーム計算の最適化技術を逐次データ同化に適用するものである。ストリーム計算では、L1 正則化を採用することにより実際上の変数選択をパラメータ推定と同時に実現している。逐次データ同化においても、状態ベクトル（ここでは時刻 t より前の時刻のパラメータベクトルに相当）のごく一部の要素でもって観測ベクトルを近似表現する、いわゆる予測エミュレータの開発研究がこの 2, 3 年、大変注目されている。我々は、データ同化の通常の観測モデルに加えて、パラメータベクトルに対してスパース性の拘束を加える擬観測モデルを導入し最適化技術を採用することで、全く新しい予測エミュレータの開発にも挑戦する。

３．研究の方法
（１）三年間計画の概略をまず説明する。アンサンブルベースの逐次フィルタからストリーム計算への技術移転等を目論むメニューＡと、その逆のメニューＢの二つの軸で研究をすすめる。初年度はメニューを実行する上で足場固めとなる情報収集に注力する。2 年度目からはメニューＡ、メニューＢを開始する。メニューＢでは最適化アルゴリズムを統計的推測問題へ拡大する作業を行なう。アルゴリズム開発と簡単な数値実験作業がメインとなる。3 年度目は、数値実験結果の十分な検討にもとづきアルゴリズムの完成度を高めるとともに、手法の具体的な問題への適用を行なう。
（２）年度毎の計画を詳しく説明する。初年度は既存手法の調査と整理にあてる。特に、機械学習分野のストリーミング計算の最新動向を調べる。問題のタイプ（判別問題、回帰問題、…）、入力データの次元、出力データのタイプ、パラメータベクトルの次元、判別関数や回帰モデル等の出力データ表現の種類、学習器内の最適化法、細かい数値技術など、さまざま項目の観点から整理した手法の分類を行なう。また、状態空間モデルに拡大解釈可能かどうかを検討する。FOBOS(Forward Backward Splitting) や RDA(Regularized Dual Averaging)への考察はもちろん、Fused Lasso のようなオフラインタイプの最適化問題の便宜的な簡易解法の調査も行なう。国内の機械学習研究を牽引する研究者が多数集う統計的機械学習センターが研究所内に設置されているメリットを生かして、ストリーム計算の理論面からの進展について適宜情報交換を行なう。
（３）①二年度目のＡメニュー：前年度の調査研究により、学習器が状態空間表現で数理モデルとして定式化されるので、ストリーム計算でアドホックに定めていたメタパラメータ（パラメータベクトルに係わる各拘束条件の重み）を、データ適用的に決定するアルゴリズムを考案する。時変メタパラメータの推定には、まず、状態空間モデルでよく利用するハイパーパラメータ（ここではメタパラメータに相当）に対する平滑（smoothness）拘束条件を採用し、その条件も含めて学習器全体を自己組織化状態空間モデルで表現する。この状態空間モデルに対して、パラメータベクトルの推定に最適化を適用する代わりに逐次データ同化手法を適用し、そのアンサンブルから最適化による推定とは異なる特性をもつパラメータベクトルを算出する方策を検討する。つまり、パラメータベクトルの推定に、アンサンブルの諸特徴量、具体的には平均やヒストグラムのピーク値などを用いた複数の決定法を考察する。
②二年度目のＢメニュー：気象・海洋シミュレーション分野において、逐次データ同化（あるいはデータ同化操作無し）で得られた状態ベクトルの系列（プロダクトと呼ばれる）と観測データのセットに対して線形回帰解析を行い、大規模なシミュレーション計算を経ずに予測値を構成する研究が現在、注目を浴びている。特に、時刻 t より前の状態ベクトルから時刻 t の観測ベクトルを予測する操作は、予測エミュレータと呼ばれる。通常は、時刻 t-1 のパラメータベクトルを説明変数、観測ベクトル内のある一要素（スカラー変数）を目的変数とする線形予測モデルを求める。回帰係数ベクトルには、その要素がほとんどゼロであるようなスパース性を仮定することで極めて低ランクの回帰係数ベクトルを求める。
③一方我々は、予測エミュレータにストリーム計算を適用することを考える。学習式の第一項に相当する観測モデルとして、元々の観測モデルの一部、つまり低ランクの観測モデルを採用する。学習器内の平滑化事前分布はそのままとし、パラメータベクトルに対してストリーム計算同様のスパース性の拘束を加える。2 年度目は、このような非線形予測エミュレータの計算アルゴリズムの開発に取り組む。
（４）①三年度目のＡメニュー：説明変数ベクトルを時刻 t の入力画像データ、観測ベクトルをあるイメージ（例えば、映像内の不審者の有無）、またパラメータベクトルをその判別器に含まれるパラメータとする時変判別問題に対して、前年度開発したアンサンブルベースの非線形フィルタを適用する。最初は機械学習コミュニティがテスト問題として利用するサンプルデータを使う。アンサンブルの諸特徴量を用いた、異なるパラメータベクトルの推定法の性能評価を、各々の判別結果およびストリーム計算の判別結果と比較する数値実験を行なう。
②三年度目のＢメニュー：　比較的低次元の状態空間モデルを用いた双子実験により、提案する非線形予測エミュレータの性能を調べる。双子実験とは、既知のシミュレーションモデルから生成したデータに、性質が既知の観測ノイズを加えて作成した観測データ

から、データ同化を用いて状態ベクトルやパラメータを推定し、どの程度真の値が復元されるかを調べる実験である。利用するモデルとしては、Kitagawa 非線形問題、Lorenz96 モデルおよびマウス概日周期モデルを想定している。
③これまで開発してきた非線形フィルタの各種情報を非線形フィルタへの適用事例とともに、論文や国際会議等での発表によりその周知につとめる。

４．研究成果
（１）まず年度毎の成果を詳しく説明する。初年度は既存手法の調査と整理にあてた。機械学習分野のストリーム計算の最新動向を調べ、問題のタイプ（判別問題、回帰問題、…）、入力データの次元、出力データのタイプ、パラメータベクトルの次元、判別関数や回帰モデル等の出力データ表現の種類、学習器内の最適化法、細かい数値技術など、複数項目の観点から整理した手法の分類を行なった。機械学習コミュニティで研究されてきた FOBOS や RDA への考察はもちろん、Fused Lasso のようなオフラインタイプの最適化問題の便宜的な簡易解法の調査も行なった。研究所のデータ同化研究開発センターのメンバーとも情報交換し、逐次データ同化手法、特にアンサンブルベースの逐次フィルタ研究の最前線を調査した。アンサンブル変換カルマンフィルタおよび局所アンサンブル変換カルマンフィルタのコンピュータ上への実装を開始した。各メンバーの調査結果をとりまとめ招待レビュー記事として発表を行なうとともに、学会でのチュートリアルセミナーも企画および実施した。
（２）3 年研究計画の課題の中間年にあたる二年度は、逐次データ同化手法のストリーム計算への適用可能性を探った。特に、ストリーム計算でアドホックに定めていたメタパラメータ（パラメータベクトルに係わる各拘束条件の重み）を、データ適用的に決定する方策ついて検討した。具体的には、時変メタパラメータの推定に状態空間モデルでよく利用する、ハイパーパラメータ（ここではメタパラメータに相当）に対する平滑（smoothness）拘束条件を採用した。その拘束条件も含めて学習器全体を自己組織化状態空間モデルで表現し、この状態空間モデルに対して逐次データ同化手法を適用した。逐次データ同化のアンサンブルから、最適化による推定とは異なる特性をもつパラメータベクトルを算出するアルゴリズムを開発した。
（３）最終年度では、二つの研究軸（A メニューと B メニュー）各々でまとめの作業に注力した。A メニューは、逐次データ同化手法のストリーム計算への適用可能性を探るものであり、逆に B メニューはストリーム計算で利用される最適化技法の逐次データ同化手法への適用を模索するものである。A メニューでは、入力ベクトルとして時刻 t の入力画像データ、観測データを二値（例えば、映像内の不審者の有無）、（潜在変数）状態ベクトルとして判別器に含まれるパラメータとする、時変判別器構築問題に対して、前年度考案したアンサンブルベースの非線形フィルタの適用を試みた。B メニューでは、比較的低次元の状態空間モデルを用いた双子実験により、ストリーム計算を模倣する非線形予測フィルタの MAP（Maximum a posteriori）解としての数値的振る舞いを考察した。さらに、これまで開発してきた非線形フィルタのソース等を Python ライブラリとして整理したプラットフォームを構築した。あわせて、論文や国際会議等での発表情報とあわせてホームページで情報を公開した。
（３）本研究成果の国内外の位置づけについて概略を述べる。機械学習のストリーム計算と逐次データ同化法の類似点に注目し、両者の優れた点を活用する方策を探る研究はそれまで皆無であった。我々は、逐次データ同化の研究において、当該分野で国内初となる教科書の出版や、国際会議での特別セッションの継続的開催などの研究実績により、逐次データ同化の分野を牽引していると国内外に認知されている。また研究代表者の樋口は、機械学習分野で国内最大研究者コミュニティである研究会 IBISML の副委員長を務めた経験があり、ストリーム計算に関する研究の最新動向にも明るい。このように、機械学習のストリーム計算と逐次データ同化法に精通している研究チームによる本研究成果は大変ユニークなものである。
（４）今後の展望について概説する。判別問題を中心とした最適化問題を逐次データ同化問題に拡大することにより、多様なビッグデータが取り扱えるようになり、結果としてストリーム計算の応用テーマの幅が格段に広がるであろう。また、機械学習分野の技術の輸入により、これまでにないタイプのエミュレータの開発ができ、非線形性の高いシミュレーションのアンサンブル予測が実現できる可能性がある。機械学習と地球科学、ビジネスとサイエンス、最適化と統計的推測、支配方程式の有無など、対照的な二つの研究分野のツールをつなぐことにより、新しい手法を開発するだけでなく、研究成果の発表等を通して研究者コミュニティの部分融合も促せたのではないかと自負している。

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 14 件）

① R. Niwayama, H. Nagao, T. S. Kitajima, L. Hufnagel, K. Shinohara, T. Higuchi, T. Ishikawa, A. Kimura, Bayesian inference of forces causing cytoplasmic streaming in

Caenorhabditis elegans embryos and mouse oocytes, PLoS ONE, Vol. **11**, No. 7,e0159917,doi:10.1371/journal.pone.0159917, 2016.査読有

② A. Sudo, T. Kashiyama, T. Yabe, H. Kanasugi, X. Song, T. Higuchi, S. Nakano, M. Saito, Y. Sekimoto, Particle filter for real-time human mobility prediction following unprecedented disaster,Proceedings of ACMSIGSPATIAL, doi:10.1145/2996913.29970000, 2016.査読有

③ S. Nakano, K.Ito, K. Suzuki, G. Ueno, Decadal-scale meridional shift of the typhoon recurvature latitude over five decades, International Journal of Climatology, Vol.36, 3819-3827, doi:10.1002/joc.4595, 2016.査読有

④ 樋口知之，人工知能はみようみまねマシンの究極形，情報管理，Vol.59，No.5，331-335,doi:http://doi.org/10.1241/johokanri.59.331，2016.査読無

⑤ S. Nakano, K. Suzuki, K. Kawamura, F. Parrenin, T. Higuchi, A sequential Bayesian approach for the estimation of the age-depth relationship of the Dome Fuji ice core, Nonlin. Processes Geophys., Vol. 23, 31-44, doi:10.5194/npg-23-31-2016, 2016. 査読有

⑥ T. Goto, Y. Hanatsuka, T. Higuchi, T. Matsui, Road condition classification using a new global alignment Kernel, Proceedings of the 2015 IEEESignal Processing Society Workshop (MLSP2015),1-6,DOI:10.1109/MLSP.2015.7324381, 2015. 査読有

⑦ S. Nakano, M.-C. Fok, P. C. Brandt, T. Higuchi, Estimation of the helium ion density distribution in the plasmasphere based on a single IMAGE/EUV image, J. Geophys. Res., Vol. 119, Issue5, 3724-3740, DOI: 10.1002/2013JA019733, 2014. 査読有

⑧ S. Nakano, M.-C. Fok, P. C. Brandt, T. Higuchi, Estimation of temporal evolution of the helium plasmasphere based on a sequence of IMAGE/EUV images, J. Geophys. Res., Vol. 119, Issue5, 3708-3723, DOI: 10.1002/2013JA019734, 2014. 査読有

⑨ 中野慎也，樋口知之，地球科学におけるシミュレーションとビッグデーター データ同化とエミュレーション－，電子情報通信学会誌，Vol. 97, No.10. 869-875, 2014.査読無

⑩ 樋口知之，統計数理の誕生とその広がり，横幹，Vol.8，No.1，14-21，2014.査読無

〔学会発表〕（計 30 件）

① A. Sudo, T. Kashiyama, T. Yabe, H. Kanasugi, X. Song, T. Higuchi, S. Nakano, M. Saito, Y. Sekimoto, Particle filter for real-time human mobility prediction following unprecedented disaster, 2016 年 11 月 1 日, San Francisco Airport Marriott Waterfront (California, USA.)

② 樋口知之，実世界データの機械学習，センサエキスポジャパン 2016－SICE 計測部門・システムインテグレーション部門共催セミナー，2016 年 9 月 29 日，東京ビッグサイト（東京都江東区）

③ 中野慎也，一般化状態空間モデルによる人工衛星磁場データの解析，2016 年度 統計関連学会連合大会，2016 年 9 月 4 日-7 日,金沢大学角間キャンパス(石川県金沢市)

④ S. Nakano, A Gaussian process model for representing the uncertainty and long-term change in typhoon behavior and its application , 2016 年 7 月 18 日-22 日, the 26th Annual Conference of the International Environmetrics Society (Edinburgh, UK)

⑤ S. Nakano, Y. Ogawa, A quantification method for the properties of diffuse and pulsating aurorae based on auroral image data 2016 年 5 月 22 日-26 日，日本地球惑星科学連合 2016 年大会(幕張メッセ，千葉県千葉市)

⑥ S. Nakano, K. Ito, K. Suzuki, G. Ueno, Decadal-scale variation of the typhoon recurvature latitude, 2016 年 5 月 22 日-26 日，日本地球惑星科学連合 2016 年大会（幕張メッセ，千葉県千葉市）

⑦ S. Nakano, P. C. Brandt, M.-C. Fok, A pilot study for reconstruction of the inner-magnetosphere by data assimilation of global ENA and EUV measurements, 2016 年 5 月 22 日-26 日，日本地球惑星科学連合 2016 年大会(幕張メッセ，千葉県千葉市)

⑧ T. Higuchi, Data Assimilation: Challenge for Big Data through Numerical Simulation , 2015 Taiwan-Japan Joint Workshop on Inverse Problems, 2015年11月22日, Department of Mathematics, National Taiwan University (Taipei , Taiwan)

⑨ T. Goto, Y. Hanatsuka, T. Higuchi, T. Matsui,Road condition classification using a new global alignment Kerne, 2015 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2015 年 9 月 19 日, Northeastern University (Boston, MA 02115 U.S.A.)

⑩ T. Higuchi, Simulation, Data Assimilation and Emulation, Asia

Oceania Geosciences Society 2015 (AOGS), 2015 年 8 月 5 日, Suntec Singapore Convention & Exhibition Centre (Suntec City, Singapore)

⑪ T. Higuchi, Stream computing and emulation in a world of the edge heavy data, International Conference for Mathematics, Statistics and Financial Mathematics (ICMSFM2014) with IASC-ARS Sessions, 2014 年 11 月 18 日, Sunway Resort Hotel&Spa (Petaling Jaya, Selangor,Malaysia)

⑫ T. Higuchi, Big data and personalization technology: Imputation, Linkage, and Stream computing, 東北大学知のフォーラム国際会議, 2014 年 11 月 6 日, 東北大学 (宮城県仙台市)

⑬ 樋口知之, 木を見て森も見るビッグデータ解析技術, 第 68 回 NHK 技研公開 2014, 2014 年 5 月 29 日, NHK 放送技術研究所(東京都世田谷区)

〔図書〕(計 0 件)

〔産業財産権〕

○出願状況 (計 0 件)
○取得状況 (計 0 件)

〔その他〕
http://www.ism.ac.jp/
http://www.ism.ac.jp/~higuchi/
http://researchmap.jp/matrix

電子情報通信学会東京支部シンポジウム
www.ieice.org/tokyo/sinpo20150305.pdf

ソフトウェア：Python 用並列粒子フィルタ
http://daweb.ism.ac.jp/support/software/P-cubed/P-cubed.html

６．研究組織
(1)研究代表者
樋口　知之 (HIGUCHI, Tomoyuki)
統計数理研究所・所長
研究者番号：70202273

(2)研究分担者
中野　慎也 (NAKANO, Shinya)
統計数理研究所・モデリング研究系・准教授
研究者番号：40378576

有吉　雄哉 (ARIYOSHI, Yuya)
統計数理研究所・データ同化研究開発センター・特任研究員
研究者番号：80735019