様式C-19

# 科学研究費補助金研究成果報告書

成 22 年 5 月 19 日現在

研究種目：若手研究（スタートアップ）
研究期間：2008～2009
課題番号：20800019
研究課題名（和文）　大規模データのための変化検出アルゴリズムとその計算アーキテクチャ

研究課題名（英文）Algorithms and computational-architectures for change detections in large-scale data

研究代表者
河原　吉伸（KAWAHARA　YOSHINOBU）
大阪大学・産業科学研究所・助教
研究者番号：00514796

研究成果の概要（和文）：本研究は，得られた時系列データを用いて，データを生成するシステム内の挙動の変化をとらえる変化検出問題に関するものである．特に，データが持つ大規模性を明示的に考慮した変化検出のアルゴリズムの導出と，更にその実装面からも考察し，効率的に実行可能な総合的な枠組みの構築を行う事を目的としたものである．本研究では，効率的に実行可能なアルゴリズムとして，近年提案された密度比推定に基づいた方法の導出を行った．また，このアルゴリズムのオンラインによる実装方法についても導出し，効率的に実行可能な枠組みを提案している．変化検出は，工学システムにおける異常診断など，極めて応用性の高い技術であるため，本研究で得られた成果は，こういった応用のための基礎的技術としても極めて重要であると考えられる．

研究成果の概要（英文）： In this research, we study the change detection problem, i.e., the problem of detecting systematic changes in data generating processes only from observed time-series data. Especially, we develop efficient algorithms that explicitly take the scale of data into consideration. To this end, we developed an algorithm based on recently proposed density-ratio estimation techniques. And, we derived an online implementation of this algorithm, which make the change detection algorithm much more efficient. Since change detection is a technology well suited for applications including fault diagnosis in engineering systems, our results would be significant also as a fundamental technology for such applications.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| ２００８年度 | 1,310,000 | 393,000 | 1,703,000 |
| ２００９年度 | 1,180,000 | 354,000 | 1,534,000 |
| 総　計 | 2,490,000 | 747,000 | 3,237,000 |

研究分野：データマイニング・機械学習
科研費の分科・細目：情報学・知能情報学
キーワード：システム工学，機械学習，データマイニング

## １．研究開始当初の背景

変化検出問題は，従来から統計分野を中心に盛んに議論されてきた重要な問題の一つである．研究代表者はこれまで，人工衛星や工場プラントなどの工学システムの異常診断を目的として，システム制御理論に基づいてこの問題について取り組んできた経緯があった．そこでは，部分空間同定法と呼ばれるシステム同定理論に基づいた変化検出アルゴリズムを導出し，その異常診断への適用に関して議論してきた．しかし大規模なデータへの適用においては，計算コストの増大や性能の低下が生じるという問題があった．

一方で，機械学習やデータマイニングなどの分野では，近年その応用分野の広がりから，適用されるデータがより大規模になった事に加え，種々の学習やパターン認識のアルゴリズムが組み合わされて使われるようになったため，アルゴリズム自体の計算量を如何に削減するかが重要な焦点の一つとなっている．特に，変化検出は実時間で実行する必要があるため，計算コストの増大はその実用可能性の点で極めて重要な問題である一方，異常検知などの多くの応用場面では，問題のクリティカル性から高い正解率が必要とされるため，その計算コストの削減のために，実行に伴う精度を極端に犠牲にする事ができない．

このような経緯と学術的背景から，近年の多様化する問題へ適用可能な変化検出の実現には，データが持つ大規模性を明示的に考慮したアルゴリズムの構築と，それを実装するための計算アーキテクチャを含めた，総合的な変化検出の枠組みが必要であるとの認識に至った事が，本研究の提案に至った最大の理由であった．

## ２．研究の目的

上記背景から，大規模なデータのための変化検出のためには，アルゴリズムとその実装の両側面から，データの大規模性を考慮した効率的な変化検出のための枠組みが必要であると考えられる．従って本研究では，近年機械学習やデータマイニング分野で議論される，大量データを扱うための方法論の視点から変化検出問題を捉え直す事により，大規模データにおける変化検出のための理論的枠組みの確立と，その実用可能性の検証を目的とする．

本研究では特に，次の２つの点を考慮した効率的な変化検出アルゴリズムの導出と検証を行う．

1. 逐次実行可能なアルゴリズムの開発

大規模データへの適用時に，メモリと計算コストを節約する最も有効な手段の一つは，アルゴリズムのオンライン化，つまり，新しくデータが入ってくる度に，前時刻までの結果を逐次的に更新可能となるようにアルゴリズムを設計する事が挙げられる．変化検出においては，何らかの意味で過去データと新しいデータを比較する必要があるが，その参照される過去データをすべて保持するのではなく，それを重点的に選択しつつ変化検出を実行するアルゴリズムの構築を行う．

2. データ内構造の利用

人工知能等の分野でオッカムの剃刀としても知られるように，必要以上に複雑なモデルを仮定する事は，タスクの性能低下へとつながる．同様の考えから，変化検出において大量データを扱う際にも，十分な精度が達成できる程度に，より簡潔なモデルを仮定して変化検出を実行する事が重要となる．そこで本研究では，機械学習分野で議論される確率グラフの学習の枠組みを利用し，データ内の確率的な独立性をデータから暗黙的に推定しつつ，変化検出を実行する方法について研究する．

## ３．研究の方法

本研究では，特に次の２つの点について研究を行った．

(1) 逐次密度比推定に基づく変化検出

本研究では，変化点検知問題を確率分布の比により変化度を計算するように定式化し，近年機械学習分野において提案された密度比推定に基づいたアルゴリズムを導出する．さらに，本アルゴリズムをオンラインで実行可能な形式を導出し，効率的に計算可能なアルゴリズムを導出する．導出されたアルゴリズムは，人工的なデータで検証すると共に，種々の実データに対しても適用し，その有用性を検証する．

(2) データ内構造の利用

データの各次元の因果関係など，データ内構造を変化検出に利用する事により，より高精度かつ効率的なアルゴリズムの導出が可能であると考えられる．本研究では，その基礎技術して，変数間の因果関係を推定する方法として，データの非正規性に基づく方法について研究を行う．特に，時系列データにおける各次元間の因果推定を推定する方法に関して検討を行い，変化検出への利用について議論する．

４．研究成果

上記で述べた２課題について各々述べる．

(1) 逐次密度比推定に基づく変化検出

ここでは，変化検出をある時点を境とした前後の一定時区間のデータを生成しているモデル(確率分布)が同じであるのか？あるいは違うものなのか？をデータのみから逐次推定する問題として定式化し，アルゴリズムの導出を行った（図１参照）．このような問題は，統計分野においては変化点検知問題として知られ，理論的枠組みの構築が古くから行われている．本研究では，近年機械学習分野で提案された密度比推定と呼ばれる原理に基づき，これを変化検出へと適用する事により，原理的に性能が保証され，かつ実用的に高い精度を持った方法の構築を行った．

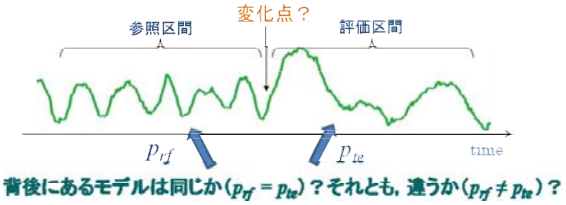


図１．変化点検知の概念図

前述のように，変化点検知問題は，２つのデータを生成する分布を比較する問題としてとらえる事ができる．今，プロセスから得られた時系列データを

$$\mathbf{Y}(t)=[y(t)^T, y(t+1)^T, ..., y(t+k-1)^T]^T$$

のように表す．このとき，ここでは，変化点検知問題を次のような２つの仮説のどちらがより確かかを比較する問題として考える（現時刻を $t$ として，対象とする時区間は $t_{rf} \le i \le t$ ）：

$$H_0: \quad p(\mathbf{Y}(i)) = p_{rf}(\mathbf{Y}(i)) \quad \text{for} \quad t_{rf} \le i \le t$$

$$H_1: \quad p(\mathbf{Y}(i)) = p_{rf}(\mathbf{Y}(i)) \quad \text{for} \quad t_{rf} \le i \le t_{te}$$
$$p(\mathbf{Y}(i)) = p_{te}(\mathbf{Y}(i)) \quad \text{for} \quad t_{te} \le i \le t$$

つまり，$H_0$ は全時区間でモデル(確率密度分布)が $p_{rf}$ であり，$H_1$ は変化点かどうかを調べたい時刻を境にモデルが $p_{rf}$ から $p_{te}$ へと変化するという仮説である．このどちらの仮説がより確からしいかは，その尤度比を計算する事で行われる．

$$\Lambda = \frac{\Pi_{i=1}^{n_{rf}} p_{rf}(\mathbf{Y}_{rf}(i)) \Pi_{i=1}^{n_{te}} p_{te}(\mathbf{Y}_{te}(i))}{\Pi_{i=1}^{n_{rf}} p_{rf}(\mathbf{Y}_{rf}(i)) \Pi_{i=1}^{n_{te}} p_{rf}(\mathbf{Y}_{te}(i))}$$

上式から分かるように，この尤度比の計算は，２つの確率密度分布 $p_{rf}$ と $p_{te}$ の比(密度比)

$$w(\mathbf{Y}) = p_{te}(\mathbf{Y}) / p_{rf}(\mathbf{Y})$$

のみに依存している事が分かる．従って，確率分布は未知な訳であるが，この比さえ計算できれば上記の判定が可能となる．つまり最終的には，各時刻において，$w$ を尤度比の式に代入して得られる次の指標を，データから計算された密度比を用いて評価する事により変化点検知を実行する事が可能となる．

$$S = \sum_{i=1}^{n_{te}} \ln \frac{p_{te}(\mathbf{Y}(i))}{p_{rf}(\mathbf{Y}(i))}$$

図２は，人工データを用いて，この指標を計算した時の例である．上図は与えたデータで，時刻 1000 の時点で，データを生成するパラメータを変更して人工的に変化点を作っている．下図は，計算された指標 $S$ で，実際の変化点の直後から値が急激に大きくなっている事が分かる．一定の閾値を与えてこの増加を検知する事により，データを生成している構造の変化をとらえる事が可能となる．

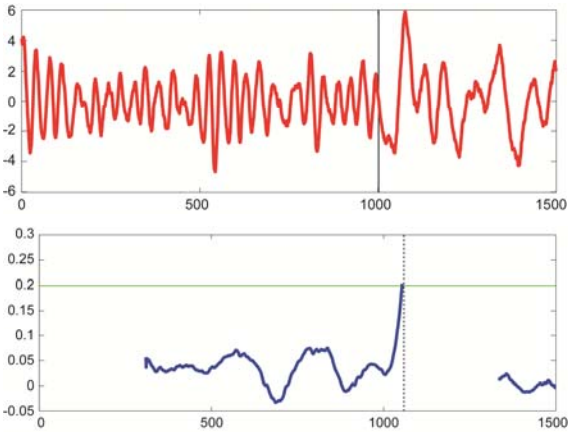


図２．変化点検知の例

密度比 $w$ の計算は，近年提案された直接推定の枠組みに基づく事で可能となる．この枠組みでは，密度比 $w$ を次式のようなカーネル関数の線形和として定義し，データからパラメータ(重み)を推定する．

$$w = \sum_{i=1}^{n_{te}} \alpha_l K_\sigma(\mathbf{Y}, \mathbf{Y}_{te}(l))$$

ここで，$K_\sigma$ はカーネル関数であり，例えば次のように定義される．

$$K_\sigma(Y,Y) = \exp(-\frac{\| Y-Y \|^2}{2\sigma^2})$$

しかし従来提案された密度比推定の方法は，バッチ的な計算(データを一括で与える必要がある)を必要とし，リアルタイムで実行する必要がある異常診断においては計算が困難となる．そこで本研究では，各時刻で新しいデータが得られる度に，パラメータ $\boldsymbol{\alpha}$ を更新していく事が可能な逐次型のアルゴリズムを導出した．図３は その疑似コードであり，図中，$n_{rf}$ および $n_{te}$ は各々参照区間及び評価区間にあるデータのサンプル数，$\eta$ や $\lambda$ はパラメータの更新率を定めるパラメータとなっている．

導出したアルゴリズムは，人工データ，及

び，実データを用いた検証により，その有用

**input**: New sample $y(t)$, the previous estimate of parameters $\boldsymbol{\alpha}$ and forgetting factors $\eta$ and $\lambda$.

1 Create the new sequence sample $\boldsymbol{Y}_{\rm te}(n_{\rm te}+1)$.

2 Update the parameters $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} \leftarrow \begin{pmatrix} (1-\eta\lambda)\alpha_2 \\ (1-\eta\lambda)\alpha_3 \\ \vdots \\ (1-\eta\lambda)\alpha_{n_{\rm te}} \\ \eta/c \end{pmatrix},$$

where $c = \sum_{l=1}^{n_{\rm te}} \alpha_l K_\sigma(\boldsymbol{Y}_{\rm te}(n_{\rm te}+1), \boldsymbol{Y}_{\rm te}(l))$.

3 Perform feasibility satisfaction:

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + (1 - \boldsymbol{b}^T\boldsymbol{\alpha})\boldsymbol{b}/(\boldsymbol{b}^T\boldsymbol{b}),$$
$$\boldsymbol{\alpha} \leftarrow \max(0, \boldsymbol{\alpha}),$$
$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}/(\boldsymbol{b}^T\boldsymbol{\alpha}),$$

where $b_l = \frac{1}{n_{\rm rf}} \sum_{i=1}^{n_{\rm rf}} K_\sigma(\boldsymbol{Y}_{\rm rf}(i), \boldsymbol{Y}_{\rm te}(l))$ for $l = 1, \ldots, n_{\rm rf}$.

4 Update as $\boldsymbol{Y}_{\rm rf}(n_{\rm rf}+1) \leftarrow \boldsymbol{Y}_{\rm te}(1)$.

図３．逐次密度比推定の疑似コード

性の確認を行った．例えば図４は，提案手法（図中 KLIEP）と，従来手法（自己回帰モデルに基づいた方法（AR），及び，カーネル密度推定に基づいた方法（KDE））を，音声データに対して適用した時の誤警報率と正解率のプロット例である．図中，左上に曲線がある方が，より低い誤警報率で高い正解率を表す．図から分かるように，提案手法は他と比べて高い性能を得ている事が分かる．本図は，特定のパラメータの場合であるが，他の場合もほぼ同様の傾向であった．本節で述べた内容は，主にデータマイニングに関する高レベルの国際会議の一つである SDM においても採録され，高い評価を得ている（業績３）．

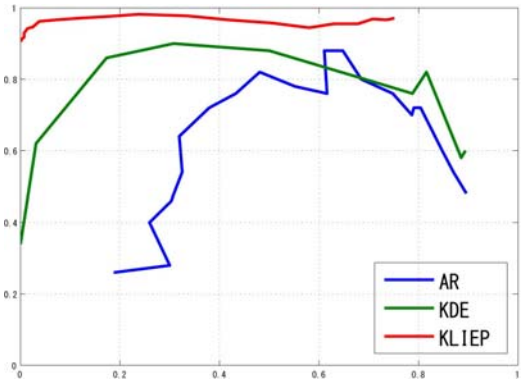


図４．誤警報率 vs. 正解率のプロット例

(2) データ内構造の利用

データ内構造をデータから推定する方法の基礎的研究として，変数間の因果関係を推定する方法に関して研究を行った．特に最近提案されたデータの非正規性を利用した方法（例えば業績２）に基づき，時系列中の変数間の因果関係の推定を中心に議論した．時系列データを $\mathbf{y}(t) := (y_1(t), \ldots, y_p(t))^T$ と表したとき，観測データから，各観測変数間の原因と結果の関係 $y_i \Rightarrow y_j$ を推定する問題についての検討を行った．

本研究では，一般的な時系列モデルの一つである，自己回帰移動平均(ARMA)モデルと呼ばれるモデルに基づいた方法を提案した．

$$\mathbf{y}(t) = \sum_{i=1}^{p} \Phi_i \mathbf{y}(t-i) + \boldsymbol{\varepsilon}(t) - \sum_{j=1}^{q} \Theta_i \boldsymbol{\varepsilon}(t-j)$$

$\boldsymbol{\varepsilon}(t)$ はノイズであり，$\Phi_i$ および $\Theta_j$ はモデルのパラメータである．本研究では，このモデルに基づき，更にデータの非正規性を用いて，変数間の因果関係を推定する方法を導出した．データが持つ非正規性は，近年因果関係の推定において極めて有用である事が発見され，この非正規性を用いて非循環型の有向グラフを推定する方法は LiNGAM 法として知られている．LiNGAM 法では，データが持つ非正規性を利用して，独立成分分析(ICA)を適用する事により変数間の関係を推定する．本研究ではこの LiNGAM 法を，ARMA に基づく時系列変数間の関係の推定に適用する．この際重要な事は，時系列においては，２種類の時間的な相関を考慮する必要があるという点である．つまり，一般にデータを取得する時間間隔は一定であるが，それに対して，より高速な(瞬間的な)影響と低速な(時間ラグのある)影響を考慮する必要がある．ARMA モデルは後者のみしか表現できないので，ここでは前者も含む次のようなモデルを考える．

$$\mathbf{y}(t) = \sum_{i=0}^{p} \Psi_i \mathbf{y}(t-i) + \mathbf{e}(t) - \sum_{j=1}^{q} \Omega_i \mathbf{e}(t-j)$$

ここでノイズ $\mathbf{e}(t)$ は非正規であると仮定し，$\Psi_i$ および $\Omega_j$ はモデルのパラメータである．ARMA モデル(a)との違いは，瞬間的な項（$i=0$）を考慮するという点のみである．このモデルの推定は，まず(1)データを用いて通常の ARMA モデル(a)を推定する．そして，(2)得られたノイズの系列 $\boldsymbol{\varepsilon}(t)$ に対して LiNGAM 法を適用する．最終的に，上の２つの式の関係からモデル(b)のパラメータの推定が可能となる．

図５は，ある物理システムから得られたデータ(4 次元)に対して，導出した ARMA-LiNGAM 法を適用した例である．瞬間的には，変数１から３，２から４の影響が存在し，そして特に変数１から４への大きな低速の影響が存在する事が推定されている．変化検出において，このような時系列の変数間の相関を推定する方法を組み込むことにより，上述のようにデータ内構造を利用した効率的・高性能な方法論の導出へとつながると考えられる．

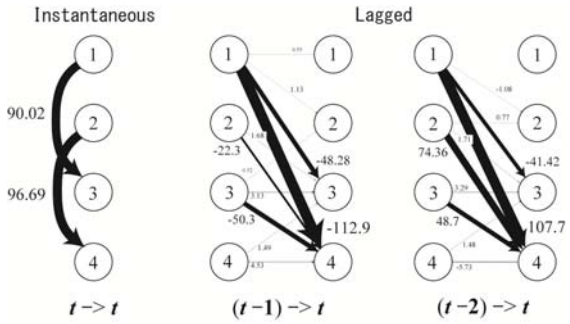

図５.ARMA-LiNGAM 法で推定した変数間の関係

５．主な発表論文等

〔雑誌論文〕（計３件）

① Y. Kawahara, K. Nagano, K. Tsuda and J. Bilmes, Submodularity cuts and applications, *Advances in Neural Information Processing Systems*, Vol.22, pp.916-924, 2009[査読有].

② S. Shimizu, A. Hyvarinen, Y. Kawahara and T. Washio, A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model, in *Proc. of the 25th Conf. on Uncertainty in Artificial Intelligence*, pp.506-513, 2009[査読有].

③ Y. Kawahara and M. Sugiyama, Change-point detection in time-series data by direct density-ratio estimation, in *Proc. of the 2009 SIAM Int'l Conf. on Data Mining*, pp.389-400, 2009[査読有].

〔学会発表〕（計７件）

① 平下 智史, 河原 吉伸, 矢入 健久, 動的環境での離散凸最適化に基づく移動ロボットのための探査法, 第 72 回 情報処理学会全国大会, 2010 年 3 月 11 日,東京大学(東京都)

② 河原 吉伸, 永野 清仁, 津田 宏治, J. Bilmes, 劣モジュラカットとその応用, 第 12 回 情報論的学習理論ワークショップ, 2009 年 10 月 19 日, 九州大学（福岡県）

③ 吉木 明博, 河原 吉伸, 矢入 健久, 次元削減技術を用いた宇宙機テレメトリの異常検知法, 第 53 回 宇宙科学技術連合講演会, 2009 年 9 月 9 日, 京都大学（京都府）

④ 乾 稔, 矢入 健久, 河原 吉伸, 町田 和雄, 次元削減の再構成誤差を用いた異常検知手法の比較, 第 23 回 人工知能学会全国大会, 2009 年 6 月 17 日, サンポートホール高松（香川県）

⑤ M. Inui, Y. Kawahara, K. Goto, T. Yairi and K. Machida, Adaptive limit checking for spacecraft telemetry data using kernel principal component analysis, in *Trans. of The Japan Society for Aeronautical and Space Science*, Vol.7, Pf_11-Pf_16, 2009 年 6 月 5 日, 浜松（静岡県）

⑥ 永野 清仁, 河原 吉伸, 岡本 吉夫, 離散凸最適化による機械学習の諸問題へのアプローチ, 第 22 回 回路とシステム軽井沢ワークショップ, 2009 年 4 月 20 日, 軽井沢プリンスホテル（長野県）

⑦ 河原 吉伸, 杉山 将, An approach for change- point detection based on direct importance estimation, 第 11 回 情報論的学習理論ワークショップ, 2008 年 10 月 30 日, 東北大学（宮城県）

〔その他〕
ホームページ等

http://www.ar.sanken.osaka-u.ac.jp/~kawahara

６．研究組織
(1)研究代表者
河原　吉伸（KAWAHARA YOSHINOBU）
大阪大学・産業科学研究所・助教
研究者番号：00514796

(2)研究分担者
該当なし

(3)連携研究者
該当なし