Vol.2017-GN-100 No.5
Vol.2017-CDS-18 No.5
Vol.2017-DCC-15 No.5
2017/1/20

データ特性を考慮した ストリーミングセンサデータ記録手法の提案

丸島晃明[†1]　峰野博史[†2]

概要：近年，センシング技術が急速に発展しており，多数のセンサから収集された情報を利活用するシステムが開発されている．特に，高周期にセンシングを行うことで生成されるストリーミングセンサデータに対して注目が集まり，ストリーミングセンサを用いたデータマイニングへの期待が高まっている．しかし，ストリーミングセンサデータの情報量は膨大であり，記憶装置への多大な負荷を考慮した場合全てのストリーミングセンサデータを正確に記録することは困難であった．そこで本論文では，ストリーミングセンサデータのうちデータの特性が急激に変化する特性変化点に着目し，ストリーミングセンサデータを記録する手法を提案する．提案手法では，ChangeFinder によってストリーミングセンサデータの特性変化点らしさ（以降，特性変化量）を高速に算出し，その後，箱ひげ図を用いてストリーミングセンサデータの特性変化量の分布を得ることで，分布から外れたデータを特性変化量の大きいデータとして記録する．一方，特性変化量の小さいデータは，概略値算出や圧縮センシングを用いて元ストリーミングセンサデータの傾向を損なわないよう非可逆圧縮し記録容量を削減する．本データ特性を考慮したストリーミングセンサデータ記録を加速度センサの実測値に適用した結果，既存の単純な概略算出手法と比較して，特性変化が顕著なデータに対して約 80%の記録容量を削減しながら高い復元精度を得ることができ，提案手法の有効性を確認した．

1. はじめに

近年，センサ技術と無線ネットワーク技術の急速な発展により，無線センサネットワークから収集された情報を利活用するシステムが開発されている．特に，センサの低コスト化と小型化，無線ネットワークの通信容量と通信速度の向上によって，現実世界の様々なモノをインターネットに接続する Internet of Things(IoT)[1]が実現可能となった．この IoT の実現に伴って，センサを用いて現実世界の様々な情報をストリーミングセンサデータとして取得できるようになり，ストリーミングセンサデータを用いたデータマイニングへの期待が高まっている．しかし，ストリーミングセンサデータは時間と共に情報量を増幅させる特性を持ち，記録するためには膨大な記録容量の記録媒体が必要となる．一方で，膨大な記録容量を要するストリーミングセンサデータを記録する記録媒体を用意することは，要する運用コストの観点から非常に困難である．加えて，一般にストリーミングセンサデータの入来速度は非常に高速であり，既存のストレージシステムでは記録が間に合わず，データを欠落させてしまう危険性が存在する．このため，ストリーミングセンサデータを記録する場合，膨大な情報量と高速な流入速度に対して記録媒体の限界を超えさせることなく効率的に記録することが重要となる．

本論文では，ストリーミングセンサデータのうちデータの特性が急激に変化する特性変化点に着目し，ストリーミングセンサデータを記録するデータ記録手法を提案する．提案手法では，ChangeFinder によってストリーミングセンサデータの特性変化点らしさ（以降，特性変化量）を高速に算出する．その後，箱ひげ図を用いてストリーミングセンサデータの特性変化量の分布を得ることで，分布から外れたデータを特性変化量の大きいデータとして記録する．一方，特性変化量の小さいデータは，概略算出や圧縮センシングを用いて元ストリーミングセンサデータの傾向を損なわないように非可逆圧縮し記録容量とディスク I/O を削減する．これにより，記憶装置への負荷を抑えながらストリーミングセンサデータの概形と特性変化を保ったデータ記録手法の確立を目指す．

以降，本稿の構成を述べる．2 章で関連研究について述べ，3 章でデータ特性変化を考慮したストリーミングセンサデータ記録手法の実現方法について説明する．4 章で性能検証実験の結果を述べ，5 章で今後の進め方についてまとめる．

2. 関連研究

膨大な記録容量を必要とするストリーミングセンサデータを記録するための手法が多数提案されている．伝統的なストリーミングセンサデータ記録手法として，データストリーム管理システム(Data Stream Management System，以降 DSMS)[2]が存在する．DSMS はストレージを用いてストリーミングセンサデータを管理することを目的としたシステムである．DSMS として実装されたシステムの例として，STREAM[3]や PipelineDB[4]などが挙げられる．DSMS は Continuous Query[5]と呼ばれる，システム内部で継続的に処理を行うクエリを採用することによって，平均値や中央値といった概略値を常に算出し続け，ディスク I/O とデータ記録量を削減する．また，Continuous Query とは別にウィンドウ技術[6]を用いることで無限長のストリーミングセンサデータを有限長に分割し，ストレージ記録量限界の課題に対応している．この DSMS の技術は現在でもストリーミングセンサデータを対象としたデータベースシステムに採用されており，ストリーミングセンサデータを用いたマイニングの発展に寄与している．しかし，Continuous

†1 静岡大学大学院総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University
†2 静岡大学学術院情報学領域 / JST さきがけ
College of Informatics, Academic Institute, Shizuoka University / JST PRESTO

ⓒ 2017 Information Processing Society of Japan

Vol.2017-GN-100 No.5
Vol.2017-CDS-18 No.5
Vol.2017-DCC-15 No.5
2017/1/20

Query によって得られる概略値は，算出する間隔を長くするほど多くの情報が失われる．特に異常検知分野における異常値といった特異な値は出現頻度が少ないため，単純な平均値や中央値の算出では丸め込まれてしまうことが多い．この場合，後に参照した際，重要となる情報が欠損してしまうこととなり，適切な記録ができているとは言えない．また，ウィンドウ技術に関しても，一部の情報は正確に記録できるものの，大半の情報を破棄するため後から参照した際には既にデータが存在しない場合があるという課題が存在する．

また，センサデバイスとそのセンサデータを利用するアプリケーションを連携し，アプリケーションの要求に応じてストリーミングセンサデータを生成する手法も存在する[7]．この手法はアプリケーションから要求があった時のみセンサからデータを生成することでストリーミングセンサデータの流入間隔を調整する．これにより，データの流入速度と流入量を削減し，現実的なストリーミングセンサデータの記録を実現している．しかし，データの要求間隔はアプリケーションによって様々であり，連続的に要求されるとは限らないため，多くの場合記録したストリーミングセンサデータは時系列性を損なう．その結果，記録されるストリーミングセンサデータはまばらなものとなり，大半の情報を破棄するウィンドウ技術と同様，後から参照した際にデータ欠損が問題を生じさせる可能性がある．

その他，近年取り組まれている手法として，間引いて記録したストリーミングセンサデータを圧縮センシング[8]で復元する手法が提案されている[9,10]. 圧縮センシングは少ない観測データから元のデータを復元することを目的とした手法であり，復元対象データがスパースな信号として表現できる場合に元のデータを復元することができる．しかし，圧縮センシングを用いて高い圧縮性能を得ようとした際，復元データが元のデータを再現しきれず，特徴的なデータを追いきれなくなる場合が存在し，その他の手法と同様に特異な情報が欠損してしまう可能性がある．このため，特性変化点といった特徴的な情報の保持と高い圧縮性能を保ったストリーミングセンサデータ向けの記録主婦法を検討することが重要となる．

3. データ特性を考慮したストリーミングセンサデータ記録手法

3.1 概要

本研究では，データ特性変化点検出アルゴリズムであるChangeFinder[11]と箱ひげ図の外れ値検出を用いたストリーミングセンサデータ記録手法を提案する．データの特徴的な変動である特性変化を動的に検出することで，大規模かつ高頻度な記録を必要とするストリーミングセンサデータに対して高い圧縮率と特性変化を保った記録を実現する．

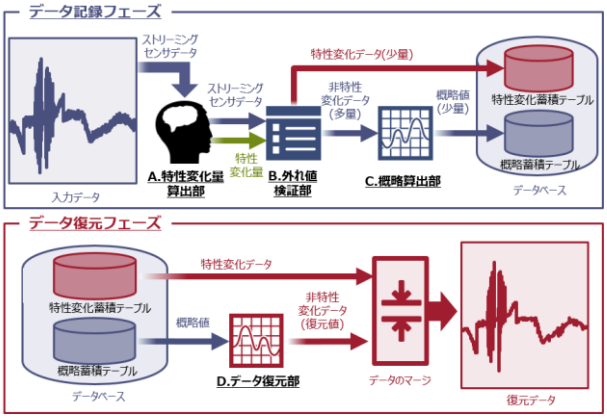


図 1：提案手法の概要

図 1 に提案手法の概要を示す．提案手法は，データ記録フェーズとデータ復元フェーズの 2 フェーズで動作する．データ記録フェーズでは，入来したストリーミングセンサデータから特性変化量算出部が特性変化量を算出する．その後，外れ値検証部が特性変化量を外れ値検証することで特性変化点を抽出しデータベースへ記録する．最後に，概略算出部が非特性変化点を概略化することで非可逆圧縮しデータベースへ記録する．データ復元フェーズでは，データ復元部がデータベースに記録した概略を用いて元ストリーミングセンサデータの概形を復元し，特性変化点とマージすることで元ストリーミングセンサデータを復元する．

3.2 特性変化量算出

この節では図 1 における A.特性変化量算出部について述べる．ストリーミングセンサデータの特性変化量の算出には時系列データにおける値の急激な変動をリアルタイムに検出する手法である ChangeFinder を用いる．ChangeFinder では AR(Auto Regression)モデルに忘却型逐次学習を導入した SDAR(Sequential Discounting Auto Regression)[12]モデルを学習に採用している．この SDAR モデルは，過去データの影響を減らした上で AR モデルのパラメータである，AR モデルの係数行列A，平均μ, 分散共分散行列 Σを推定する．式(1)に示すIを最大化するようなA, μ, Σを求めることで，時刻 t における時系列モデル(時刻 t までのデータを用いた確率密度関数)を得る．このとき，r(0<r<1)を忘却係数と呼び，1 に近いほど過去データの影響を減らして推定を行う．また，k はパラメータを推定する AR モデルの次数を指す．

$$I = \sum_{i=1}^{t} (1-r)^{t-i} \, logP(x_i | x^{i-1}, A_1, \ldots, A_k, \mu, \Sigma) \qquad (1)$$

また，SDAR モデルはモデルのパラメータ推定時に過去のパラメータと入来したデータx_tを用いて逐次的に推定を行うため，高速にモデルのパラメータを更新できる．これにより，ChangeFinder の計算量を$O(n)$に抑えることができるため，流入速度の速いストリーミングセンサデータに対し

ⓒ 2017 Information Processing Society of Japan

Vol.2017-GN-100 No.5
Vol.2017-CDS-18 No.5
Vol.2017-DCC-15 No.5
2017/1/20

て高速かつ逐次的に特性変化量の算出を行うことができる．

ChangeFinder の動作フローを**図 2** に示す．ChangeFinder はまず流入した時系列データx_tを SDAR で学習を行い，時系列データx_tに対する確率密度関数$p_t(x_t)$を得る．その後，$p_{t-1}(x_t)$を用いてx_tに対する対数損失を求め，これをx_tの外れ値らしさを表す外れ値スコア$Score(x_t)$とする．$Score(x_t)$の算出式を式(2)に示す．

$$Score(x_t) = -\log p_{t-1}(x_t|x^{t-1}) \quad (2)$$

ただし，この外れ値スコア$Score(x_t)$ではスコアリングにノイズの影響を大きく受ける．そこで，ChangeFinder では移動平均を算出することで外れ値スコアを平滑化し，ノイズの影響を減らす．平滑化スコアy_tの算出式を式(3)に示す．

$$y_t = \frac{\sum_{i=t-T}^{t-1} Score(x_t)}{T} \quad (3)$$

この平滑化した外れ値スコアy_tを用いて SDAR で再度学習し，確率密度関数$q_t(y_t)$を得る．その後$q_{t-1}(y_t)$を用いて対数損失を求めることで時系列データ中の特性変化量$Score(t)$を得る．$Score(t)$の算出式を式(4)に示す．

$$Score(t) = -\log q_{t-1}(y_t|y^{t-1}) \quad (4)$$

このように，ChangeFinder は SDAR モデルを利用した二段階学習でノイズの影響を除去しながら特性変化量を算出できる．この ChangeFinder を用いてストリーミングセンサデータ中の特性変化量を算出することで特性変化の推移を明らかにし，特性変化量に基づき特性変化点を抽出する．

3.3 特性変化点抽出

この節では図 1 における B.外れ値検証部ついて述べる．提案手法では，ChangeFinder を用いて算出した特性変化量を一次元データ集合の分布を表す箱ひげ図として表現する．これにより，ストリーミングセンサデータにおける特性変化量の分布を得る．箱ひげ図の概要を**図 3** に示す．箱ひげ図は第一四分位数から第三四分位数までの区間を箱として表現し，第一四分位数からデータ集合中の最小値，第三四分位数からデータ集合中の最大値までの区間をひげとして表現する．この時，ノイズ等の外れ値がひげの長さに影響を及ぼすため，外れ値を考慮した箱ひげ図では**図 4** のように箱の 1.5 倍の長さより箱から離れた場所に存在するデータを外れ値として判断し，その外れ値を除外した上でひげの長さを決定する．この箱ひげ図の外れ値検出を特性変化量に適用し，箱の長さの 1.5 倍より遠い距離に存在するデータを特性変化量の大きい，特性変化点として扱う．また，箱の長さの 1.5 倍以内の距離に存在するデータを特性変化量の小さい，非特性変化点として扱う．

提案手法では，ある時刻 t のストリーミングセンサデータx_tが入来した時，式(5)に示す直近 b 件の特性変化量$S(x_i)$の集合 X を利用して箱ひげ図として表現する．

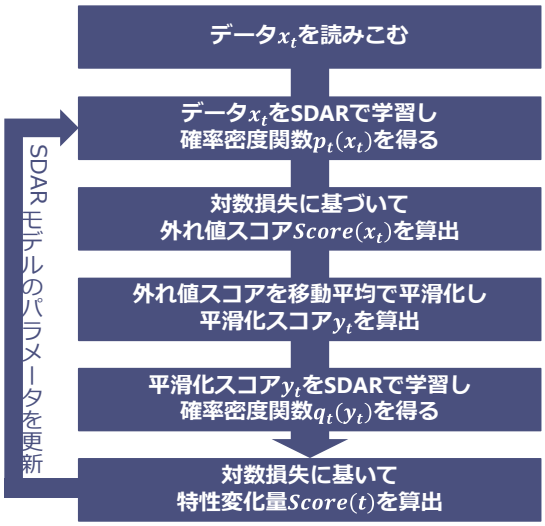


図 2：ChangeFinder の動作フロー

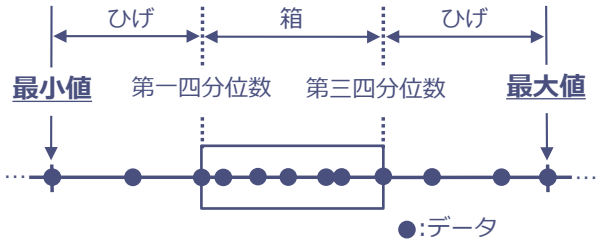


図 3：箱ひげ図の概要

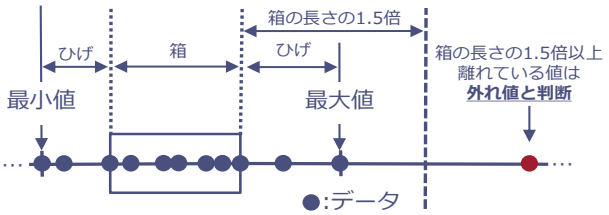


図 4：外れ値を考慮した箱ひげ図

$$X = \{ S(x_i) \mid t - b \le i \le t - 1 \} \quad (5)$$

その後，x_tから算出された特性変化量$S(x_t)$が箱から 1.5 倍より遠い距離に存在すると判断された場合，x_tを特性変化点として判断する．一方，$S(x_t)$が箱から 1.5 倍の長さ以内の距離に存在すると判断された場合，x_tは非特性変化点であると判断し，概略算出部へ渡すことで概略化することで非可逆圧縮する．

なお，箱ひげ図として表現するために利用する特性変化量のデータ範囲である b に関してはシステムパラメータとする．この b の値はx_tの特性変化量である$S(x_t)$が特異な値であるかどうかを検証するために利用する他時刻の特性変化量のデータ数を表す．このため，b が小さいほど提案手法がx_tを特性変化点であるという判断を下しやすくなる．大規模な特性変化のみ抽出したい場合は b を大きく設定し，小規模な特性変化も含めて抽出したい場合は b を小さく設定するなど，状況に合わせたパラメータ設定が可能である．

ⓒ 2017 Information Processing Society of Japan

Vol.2017-GN-100 No.5
Vol.2017-CDS-18 No.5
Vol.2017-DCC-15 No.5
2017/1/20

3.4 概略算出

この節では図 1 における C.概略算出部について述べる．3.3 節の特性変化点抽出で非特性変化点であると判断されるデータはストリーミングセンサデータ中の多数を占めることが多いと考えられる．このため，ディスク I/O や記録可能容量の限界から非特性変化点をそのままデータベースへ記録することは困難である．一方で，非特性変化点は特性変化点と比較して，概して値の変動幅が小さいという特徴を持つため，概略化したときに発生する誤差は特性変化点より小さくなることが期待できる．このため，これらの非特性変化点を概略化することで非可逆圧縮を行い，高い圧縮率を得ることでディスク I/O と記録容量を削減する．

概略化の例として，特性変化や外れ値，ノイズの影響を受けにくい中央値の算出が挙げられる．非特性変化点の中央値を算出し続けることで元ストリーミングセンサデータの傾向損失を抑えながら非特性変化点の記録量を削減することができる．その他の概略算出方法として，間引いてサンプリングしたストリーミングセンサデータをスパースな信号へ変換し，間引いた情報を圧縮センシングによって復元することも可能である．この場合，3.5 節で述べるストリーミングセンサデータ復元時に多くの計算リソースと処理時間を要するが，一般に元ストリーミングセンサデータとの誤差が小さいデータを復元することが可能である．

また，ストリーミングセンサデータの特性が変化した際，その後のデータは特性変化以前のデータと比べて値の分散や中央値などが異なる可能性が高い．このため，特性変化以前のデータと特性変化後のデータを一括りにした概略算出を行った場合，復元データが元ストリーミングセンサデータから大きく乖離する可能性がある．そこで，復元したストリーミングセンサデータと元ストリーミングセンサデータの誤差を減らすために概略の算出間隔を特性変化点が検出される都度変更し，特性変化点の検出間隔に応じて概略の算出間隔を動的に調整する．なお，特性変化点が長期間検出されない場合，ストリーミングセンサデータの値が緩やかに変化している可能性が考えられるため，ストリーミングセンサデータの傾向を保持するために定期的に概略算出を行う. 現在は暫定的に 10 回に 1 回の定期的な概略算出を行っているが，この長期間特性変化点が検出されない場合の概略算出間隔は用途と求めたいデータ圧縮性能に応じて変更が可能である．

3.5 ストリーミングセンサデータ復元

この節では，図 1 における D.データ復元部について述べる．提案手法で記録したストリーミングセンサデータを参照する際，特性変化点抽出と概略算出によってデータベースに記録した特性変化点と非特性変化点を用いて元ストリーミングセンサデータの復元を行う．データの復元は，概略として記録した非特性変化点を用いて記録しなかった非特性変化点を補間することで行う．中央値の算出で概略を生成した場合，算出した中央値を算出に要したサンプル数だけ引き伸ばすことで補間する．単純なアルゴリズムではあるが，中央値は算出に要したサンプルとの距離の総和が最も小さくなるという特徴を持ち，加えて特性変化点抽出で中央値の算出間隔を最適化していることから，中央値の引き伸ばしでもある程度近しいデータを復元できると考える．

その他，圧縮センシングを用いてストリーミングセンサデータを概略化した場合，間引いて観測されたストリーミングセンサデータを基底追跡といった圧縮センシングにおける復元手法を用いて元ストリーミングセンサデータを推定し，復元を行う．その後，復元したデータに上乗せする形で特性変化点のデータをマージし，特性変化点の情報を保ったストリーミングセンサデータの復元を行う．

4. 性能評価実験

4.1 実験内容

従来用いられることの多かった定期的な概略算出と比較して，提案手法が記録データ量を削減しながら，ストリーミングセンサデータ中の大きな特性変動を保ったデータ復元が可能であるのかを検証する．加速度センサから取得されたストリーミングセンサデータの記録と復元を実施し，データの復元性能とデータ圧縮性能として評価する．この評価実験を行うにあたり，ストリーミングセンサデータの記録量の削減方法として一般的に用いられる定期的な概略算出とデータの復元性能を比較する．なお，比較対象として算出する概略値には，提案手法と同様に中央値を利用する．このとき，提案手法のデータ圧縮性能と同程度のデータ圧縮性能となるよう, 中央値の算出間隔を適宜調整する．

図 5 に評価実験の概要を，**表 1** に評価実験の実験環境を示す．図 5(a)はストリーミングセンサデータを圧縮して記録し, 圧縮性能を評価する実験の概要を示す. また, 図 5(b)は図 5(a)の実験で記録したデータからストリーミングセンサデータを復元し, 復元性能を評価する実験の概要を示す. まず図 5(a)に示すように 3 軸加速度センサから取得した値のうち，X 軸の値をまとめた csv ファイルを用意する．この csv ファイルを用いて，加速度センサの値を WebSocket でストリーミングセンサデータとして送信する環境を構築する．その後受信したデータに対して**表 2** に示すシステムパラメータを使用して変化点抽出と概略算出を行いデータベースへ記録する．システムパラメータは，提案手法の特徴を活かせるように，目に見えて特徴的なデータを特に判断しやすいように設定する．その後，実際にデータベースへ記録したレコード数と csv ファイルのレコード数を比較し，データ圧縮性能を圧縮率として評価する．圧縮率 Compression Ratio は式(6)によって表される．R_num は csv ファイルに記録されている真値の数を示し，C_num は提案

ⓒ 2017 Information Processing Society of Japan

Vol.2017-GN-100 No.5
Vol.2017-CDS-18 No.5
Vol.2017-DCC-15 No.5
2017/1/20

手法によって抽出された特性変化点の数，O_num は概略算出を行った回数を示す．

$$\text{Compression Ratio} = \left(\frac{C\_num + O\_num}{R\_num}\right) * 100 \qquad (6)$$

その後，図 5(b)に示すようにデータベースへ記録した情報を用いてストリーミングセンサデータの復元を行う．この時，式(3)によって算出された圧縮率と同程度の圧縮率となるよう算出間隔を調整した中央値を真値から生成する．

加えて，提案手法で復元したストリーミングセンサデータと同程度の圧縮率になるよう算出間隔を調整した中央値に対して，それぞれ式(7)の最大絶対誤差（MaxAE）と式(8)の絶対平均誤差（MAE），式(9)の絶対平均誤差率（MAPE），式(10)の二乗平均平方誤差（RMSE），式(11)の正規化二乗平均平方誤差（NRMSE），式(12)の相対絶対誤差（RAE），式(13)の相対二乗誤差（RSE）の 7 つの評価指標を算出することで復元性能を評価する．以降の数式において，N はデータ数，a_iが時刻 i における真値を指し，p_iが時刻 i における復元データを指す．a_{max}は全ての真値の最大値，a_{min}は全ての真値の最小値を指し，$\bar{a}$は全ての真値の平均を指す．

$$MaxAE = \max_{1<i<N} |a_i - p_i| \qquad (7)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |a_i - p_i| \qquad (8)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \left|\frac{a_i - p_i}{a_i}\right| * 100 \quad (a_i \neq 0) \qquad (9)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (a_i - p_i)^2} \qquad (10)$$

$$NRMSE = \frac{RMSE}{a_{max} - a_{min}} \qquad (11)$$

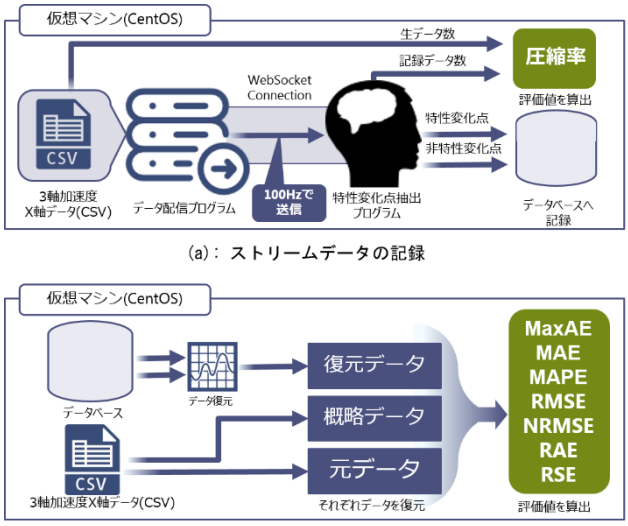
$$RAE = \frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i=1}^{n} |\bar{a} - a_i|} \qquad (12)$$

$$RSE = \frac{\sum_{i=1}^{n} (p_i - a_i)^2}{\sum_{i=1}^{n} (\bar{a} - a_i)^2} \qquad (13)$$

MaxAE は実験結果における全ての誤差のうち，絶対値が最も大きかった誤差を指す．特性変化点は一般的に値の変動幅が大きく，正しく特性変化点を復元できなかった場合 MaxAE が大きくなりやすい．このため，この値が小さいほど特性変化点を正しく復元できていると言うことができる．

MAE は絶対誤差の平均を指し，真値と復元値間で平均してどの程度の誤差が発生したかを示す．全ての誤差を平等に評価するため，出現頻度の多い非特性変化点の誤差平均に影響を及ぼしやすい．また，MAPE は絶対誤差率の平均を指す．MAPE は真値から見たパーセンテージを表すため，真値の大きさに結果が左右されやすいという特徴を持つ．

RMSE は二乗誤差平均の平方根，NRMSE は正規化した



(a)：ストリームデータの記録

(b)：ストリームデータの復元

図 5：評価実験の概要

表 1：評価実験環境

項目	値
OS	CentOS7.2.1511 (64bit)
プロセッサ	Intel® Core™ I5-4570
クロック周波数	3.20GHz
割当コア数	1
メインメモリ	4GB
データベース	MongoDB 2.6.12

表 2：使用したシステムパラメータ

項目	設定値
忘却係数 r (0<r<1)	0.001
AR モデルの次数 k	1
移動平均平滑化時の区間長	3
箱ひげ図作成時のデータ数	100

RMSE であり，それぞれ復元データが真値からどれほど乖離しているかを指す．RMSE と NRMSE は式中に二乗誤差を用いているため，誤差のバラ付きに影響を受けやすいという特徴を持つ．また，NRMSE は真値の最大値と最小値間の誤差で基準を設けているため，異なるデータ間での比較が可能である．

RAE は全ての真値の平均と真値の絶対誤差平均と復元データと真値の絶対誤差平均の比を，RSE は全ての真値の平均と真値の二乗誤差平均と復元データと真値の二乗誤差平均の比を表す．平均値のみを用いた非常に単純なモデルと誤差状況を比較してどれだけ性能が向上したかを示している．この値が 1 を下回る時，全ての真値の平均値のみを用いる単純な復元より良い復元ができていると言える．

圧縮性能の指標はどれほどの記録データを削減できたかを示すため値が大きいほど性能が良い．また，復元性能の指標は誤差量を示すため結果が 0 に近づくほど良い結果

ⓒ 2017 Information Processing Society of Japan

Vol.2017-GN-100 No.5
Vol.2017-CDS-18 No.5
Vol.2017-DCC-15 No.5
2017/1/20

となる．本評価実験では，圧縮性能を示す圧縮率と復元性能を示す 7 つの評価指標と併せて，真値と復元データの波形を比較しながら総合的に性能を評価する．

また，本評価実験に用いた実験データを**表 3** に示す．実験データは Human Activity Sensing Consortium (HASC)[13]が提供する装着型センサデータベースである HASCcorpus2014 を利用する．この HASCcorpus2014 に含まれるデータのうち，被験者 person1001 による stay（静止），walk（歩行），jog（ジョギング），skip（スキップ），stUp（階段を上る），stDown（階段を下る）の 6 動作を行った時に得られた加速度センサの計測値を用いる．stay のデータは計測値が-0.15 付近で微振動を繰り返しているデータとなる．このデータは値の範囲は小さいが，非常に振動が細かい．walk のデータは計測値が定期的に上下に揺れる動作を繰り返すデータである．jog のデータは計測値が激しく変動を続ける，標準偏差が大きいデータとなる．skip のデータは，大半は大きく変動しないデータであるが，不定期に大きな変動が発生する．また，計測値の絶対値の最大が最も大きいデータでもある．suUp のデータは一度急激に値が上昇し，その後振動しながら緩やかに値が下がっていく，という動作を繰り返すデータである．stDown のデータは，skip のデータのように大半は大きく変動せず，不定期に一部のデータが大きく変動する．また，skip のデータと比較して，大きな変動が発生した時の変動の継続時間が長いという特徴を持つ． なお，この加速度センサの計測は同一被験者によって複数回行われているが，実験にはそれぞれの動作における初回計測時のデータを利用する．

4.2 実験結果・考察

図 6 に各行動の加速度センサ計測値を用いた復元の結果を示す．図 6 のグラフにおいて，真値を黒色の実線，提案手法による復元結果を橙色の実線で表す．また，青色の点線で比較対象である同程度の圧縮率に調整した中央値を表す．なお，縦軸の単位は重力加速度 G(約$9.80665\mathrm{m}/s^2$)であり，横軸は ms 単位での時間を示す．**表 4** に各行動の加速度センサ計測値の統計量を，**表 5** に**各行動の加速度センサ計測値**における復元結果の圧縮性能と復元性能を示す．Compression Ratio が復元結果の圧縮性能を表し，MaxAE，MAE，MAPE，RMSE，NRMSE，RAE，RSE が復元性能を表す．

図 6(a)の stay では特性変化点の記録が殆ど行われなかった．これは値の振動が非常に細かく，非常に高頻度に大きな変動が出現したためである．その結果，特性変化点の出現時に算出される特性変化量が小さくなり，上手く特性変化点を抽出できなくなったと考えられる．特性変化点の記録が行われなかった結果，表 4 の stay における復元性能を見ても，提案手法による復元結果は同程度の圧縮率とした中央値とほぼ同じものとなっている．特に，MaxAE は中央値と同じ値を示している．その他の指標に関しては，11000ms～12000ms の間に存在する，微量に抽出された特性変化点の誤差軽減が影響し中央値に勝る結果となったと考えられる．一方で，変化点抽出による誤差軽減が少なかったため，他の行動データの結果と比較したとき NRMSE が大きめの値をとり，高い性能を示すことができているとは言えない．この結果から，提案手法は変動が細かく発生するデータに対しては大きな効果を発揮できず，既存の定期的な概略算出と同程度の性能となることが分かる．

表 3: 利用した HASC corpus2014 のデータ

項目	値
被験者	Person1001
利用動作データ	stay(静止) walk(歩行) jog(ジョギング) skip(スキップ) stUp(階段を上る) stDown(階段を下る)
サンプリング秒数	20 秒
サンプリングレート	100Hz

walk のデータでは，MaxAE を約ニ割程度削減し，その他の評価指標においても微量ではあるが各評価項目において復元性能が向上していることが見て取れる．復元性能の向上が微量である点については，提案手法の特性が関係していると考えられる．提案手法は特性変化点の復元性能は非常に高いが，同程度の圧縮性能の中央値と比較して非特性変化点の復元性能が低くなる．このため，平均して結果を得る評価指標では大きな性能向上が得られなかったと考察する．加えて，図 6(b)の walk のグラフを見るとマイナス方向の大きな変動を多く抽出し復元でき，大きく誤差を軽減できていることが分かる．一方で，マイナス方向の変動の後に生じるプラス方向への大きな変動を上手く抽出しきれていない．これは，特性変化抽出後にやってきたプラス方向の変動が，抽出したマイナス方向の変動と同様に大きな変動であるという特徴を持つため，特性変化点として抽出されなかったと考えられる．このように，概略算出間隔の差による誤差と，抽出しきれなかったプラス方向の変動が要因となり，結果として若干量の性能向上となったと考察する．

jog のデータを復元する時，圧縮性能が大きく落ち，それに伴い復元性能も中央値を大きく下回っていることが分かる．これは，図 6(c)に示すように，jog のデータは stay と同様に激しく変動し多くの特性変化点が生じていることに起因する．また，jog のデータは stay と比較して変動の間隔が短いため，ある程度の特性変化点を抽出している．

ⓒ 2017 Information Processing Society of Japan

Vol.2017-GN-100 No.5
Vol.2017-CDS-18 No.5
Vol.2017-DCC-15 No.5
2017/1/20

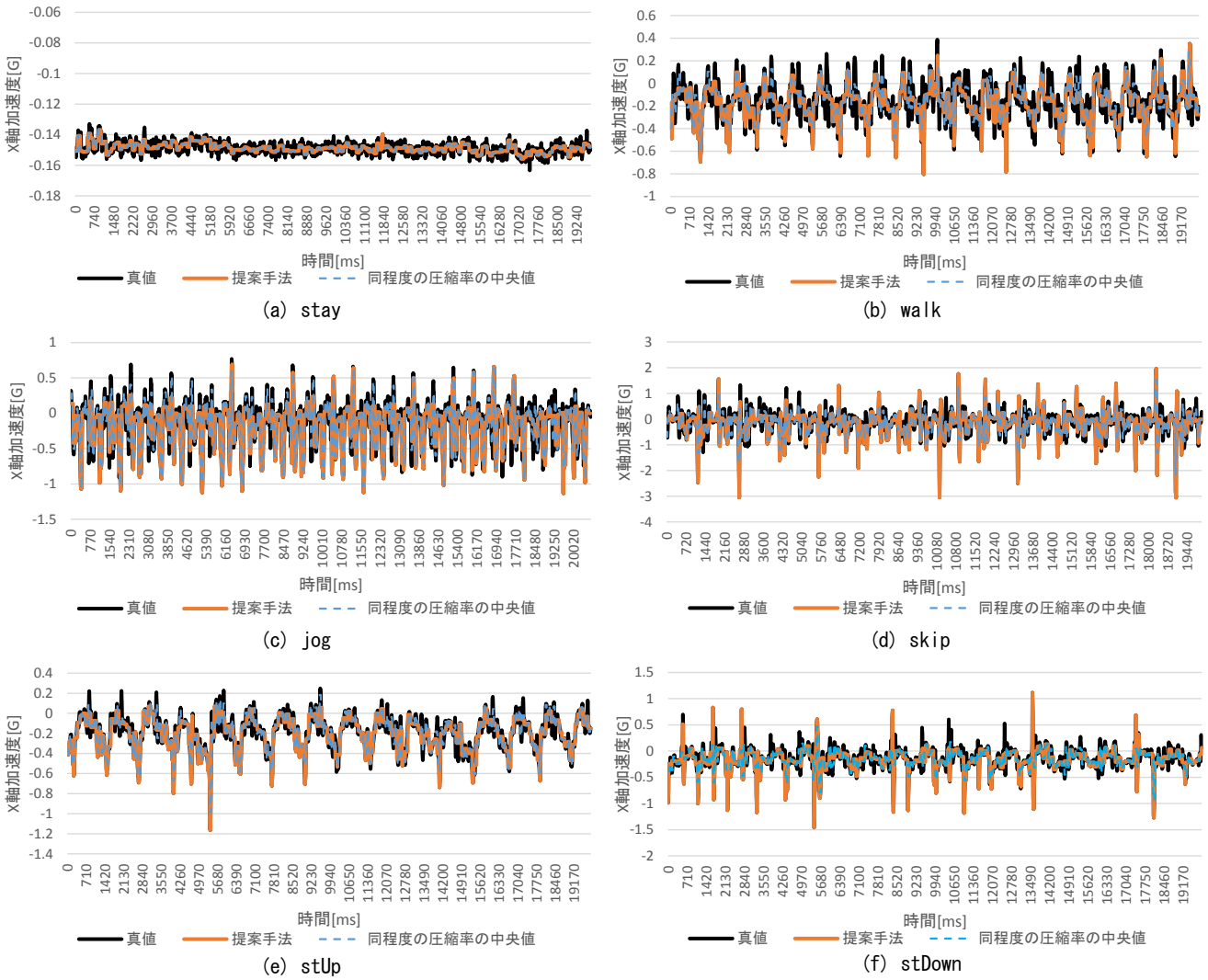


図 6: 各行動の加速度センサ計測値における復元の結果

表 4: 各行動の加速度センサ計測値の統計量

統計量	(a) stay	(b) walk	(c) jog	(d) skip	(e) stUp	(f) stDown
標本分散(G^2)	0.0000142	0.0299	0.113	0.239	0.0306	0.0499
平均値(G)	-0.149	-0.162	-0.169	-0.158	-0.196	-0.154
最大値(G)	-0.133	0.389	0.768	1.95	0.249	1.11
最小値(G)	-0.163	-0.805	-1.13	-3.06	-1.16	-1.46

表 5: 各行動の加速度センサ計測値における復元性能と圧縮性能

評価指標	(a) stay		(b) walk		(c) jog		(d) skip		(e) stUp		(f) stDown	
	中央値	提案	中央値	提案	中央値	提案	中央値	提案	中央値	提案	中央値	提案
Compression Ratio(%)	89.2		83.4		74.1		74.7		79.4		80.5	
MaxAE(G)	**0.0124**	**0.0124**	0.586	**0.486**	**0.688**	0.906	2.51	**1.83**	0.532	**0.376**	1.22	**0.603**
MAE(G)	0.00196	**0.00192**	0.0736	**0.0717**	**0.092**	0.115	**0.165**	0.172	**0.0387**	0.0493	0.0845	**0.0718**
MAPE(%)	1.32	**1.29**	179	**160**	**157**	304	**236**	262	**135**	141	**191**	197
RMSE(G)	0.00263	**0.00257**	0.109	**0.104**	**0.14**	0.183	0.283	**0.274**	**0.062**	0.0739	0.155	**0.113**
NRMSE(G)	0.0866	**0.0848**	0.091	**0.087**	**0.0739**	0.0961	0.0565	**0.0547**	**0.0439**	0.0523	0.0601	**0.044**
RAE(%)	0.671	**0.654**	0.545	**0.53**	**0.337**	0.423	**0.462**	0.479	**0.277**	0.353	0.535	**0.455**
RSE(%)	0.487	**0.467**	0.393	**0.36**	**0.174**	0.295	0.334	**0.314**	**0.125**	0.178	0.479	**0.257**

ⓒ 2017 Information Processing Society of Japan

Vol.2017-GN-100 No.5
Vol.2017-CDS-18 No.5
Vol.2017-DCC-15 No.5
2017/1/20

しかし，多くの特性変化点を抽出したため圧縮性能が低下した．このため，比較対象として用いている中央値の算出間隔が短くなり，性能が向上している．その結果，提案手法が非特性変化点の復元時に発生する誤差を特性変化点の誤差軽減でカバーしきれず，中央値と性能が開いている．提案手法は本来ストリーミングセンサデータ中に特性変化点は少ないということを前提にしているため，stay や jog のように激しく変動するデータの復元を得意としないことがこの結果から分かる．

図 6(d)の skip は，特徴的な特性変化点を概ね上手く検出できているように見える．表 4 の数値面から見ても，MaxAE を約三割軽減できており，概略算出による特性変化点の丸め込みを防ぐことができていることが分かる．また，NRMSE も stay，walk，jog と比較して小さく，全体的に誤差が小さいことが分かる．一方で，MAE は中央値と比較して低い性能を示しているが，これは非特性変化点の概略算出間隔が影響していると考えられる．しかし，特徴的な特性変化の大半を正確に再現し，その他を概形として追えており，平均的な誤差こそ中央値と比較して大きいが，提案手法の目的とするところは達成できていると考える．

図 6(e)の stUp は，多くの急激な変動を特性変化点として抽出できているが，その後の緩やかな値の降下時に上手く値を復元できていないことが多い．特に 5900ms～6500ms の間では，真値がある程度大きな幅で変動しているにも関わらず提案手法は追いきれていない．これは，大きな変動後の緩やかな値の降下時に，stay や jog のように値が分散しながら変動しているためであると考えられる．これが影響し，全体的な復元性能で見た時提案手法は中央値に劣っている．

図 6(f)の stDown は，特にマイナス方向の変動の大半を抽出することができており，MaxAE を大きく軽減できている．これにより，大きな誤差が影響しやすい RMSE や NRMSE，RSE を大きく軽減できている．しかし，こちらも walk や skip と同様にプラス方向の変動を抽出できていない．このため，提案手法の特性変化点への反応をより過敏にする必要があると考える．

以上の結果から，提案手法は値の変動が少なく，まばらに特性変化点が出現するデータに対して高い復元性能を得られることが分かる．また，今回の実験で設定したパラメータでは，特性変化の直後に続く形となる特性変化を上手く抽出できない場合があるという特性が明らかになった．この結果は特性変化点への反応が鈍感であることが原因であると考えられ，値の変動の激しさや特性変化点の出現間隔に合わせてシステムパラメータを設定し，特性変化点への反応を過敏にすることによってより復元性能が向上すると期待できる．このため，値の分散と言ったデータの特徴に応じてシステムパラメータを決定する方法を定めることが提案手法の性能向上を行う上で効果的あると考える．

5. おわりに

本論文では，ストリーミングセンサデータのうちデータの特性が急激に変化する特性変化点に着目し，ストリーミングセンサデータを記録するデータ記録手法を提案した．加速度センサの計測データを用いた評価実験の結果，激しい変動を持たないデータに対して，従来の概略算出手法と比較して最大誤差を軽減できることを示した．

今後の課題として，復元対象のデータ特性に合わせたシステムパラメータの決定方法を検討することが挙げられる．また，加速度センサの計測値を実際に復元した値を利用して行動認識を行えるかといった，復元データの実用性について評価を行う予定である．

謝辞 本研究の一部は，東北大学電気通信研究所における共同プロジェクト研究によって実施したものである．

参考文献

[1] IEEE, Towards a definition of the Internet of Things (IoT), http://iot.ieee.org/images/files/pdf/IEEE_IoT_Towards_Definition_Internet_of_Things_Revision1_27MAY15.pdf, (参照: 2016/12/12)
[2] R. Motwani, et al., Query processing, resource management, and approximation in a data stream management system, Proceedings of the 2003 CIDR Conference, pp.1-12, 2003.
[3] A. Arvind, et al. STREAM: the stanford stream data manager (demonstration description), Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pp.665-665, 2003.
[4] PipelineDB, PipelineDB—The Streaming SQL Database., https://www.pipelinedb.com/, (参照: 2016/12/12).
[5] A. Arvind, et al., The CQL continuous query language: semantic foundations and query execution, The VLDB Journal—The International Journal on Very Large Data Bases, vol.15, no.2, pp.121-142, 2006.
[6] B. Babcock, et al., Models and issues in data stream systems, Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp.1-16, 2002.
[7] S Madden, et al., TinyDB: An Acquisitional Query Processing System for Sensor Networks, ACM Transactions on database systems, vol.30, no.1, pp.122-173, 2005.
[8] E. J. Candes, et al. An introduction to compressive sampling, IEEE signal processing magazine, vol.25, no.2, pp21-30, 2008.
[9] S. Li, et al., Compressed sensing signal and data acquisition in wireless sensor networks and internet of things, IEEE Transactions on Industrial Informatics, vol.9, no.4, pp.2177-2186, 2013.
[10] M. Leinonen, et al., Sequential compressed sensing with progressive signal reconstruction in wireless sensor networks, IEEE Transactions on Wireless Communications, vol.14, no.3, pp.1622-1635, 2015.
[11] J. Takeuchi, et al., A unifying framework for detecting outliers and change points from time series, IEEE transactions on Knowledge and Data Engineering, vol.18, no.4, 2006.
[12] 山西健司, データマイニングによる異常検知, 2009.
[13] N. Kawaguchi, et al., HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings, Proceedings of the 2nd Augmented Human International Conference, pp.27, 2011.

ⓒ 2017 Information Processing Society of Japan