Vol.2012-MBL-61 No.25
Vol.2012-UBI-33 No.25
2012/3/13

同時観測クラスタ群を利用した相関に基づくマルチデータストリーム予測方式

荒井健次[†]　白石陽[‡]　高橋修[‡]

近年，様々な分野・領域でデータストリームと呼ばれる時系列データが着目されており，その未来予測を行うストリーム予測技術に注目が集まっている．ストリーム予測技術の多くは単一のデータストリームの過去の計測値の傾向を用いるが，その場合，未来の計測値が過去の傾向と異なる変化を示す場合，精度が悪化するという問題がある．しかし，他のデータストリームとの間に相関が存在する場合，相関を考慮することで予測精度が向上する可能性がある．そこで，本稿はデータストリームの状態を同時観測クラスタ群により表現し，データストリーム間の相関検出とストリーム予測を行う手法を提案し，その有効性を検証する．

A Method for Multi Stream Prediction Based on Correlation Using Simultaneous Observation Cluster Group

KENJI ARAI[†]　YOH SHIRAISHI[‡]
OSAMU TAKAHASHI[‡]

Research on one of time series data, which called data stream, has attracted a great deal of attention in many fields in recent years. Therefore, stream prediction technologies have attracted the attention of stream mining technologies. When we want to obtain the predicted value of a certain single data stream, most methods use past trend of measured data on the data stream. However, if future trend will change from past trend, the accuracy of prediction will be worsening. We think that correlations, such as synchronization, can be used for method of predicting streams, and their accuracy might be better. We suggest a method of expressing trend of data stream using simultaneous observation clusters in this paper. In addition, we suggest a method of detecting correlation using simultaneous observation cluster, and a method of prediction streams based on correlations. We also discuss our demonstration of the efficacy of these methods

1. はじめに

近年，センサネットワークや株価データを初めとした様々な分野・領域でデータストリームが着目されている．データストリームは時系列データとして表され，その特徴として，リアルタイムにデータ長が増加する無限長のデータであることが挙げられる．このようなデータストリームから有用な規則やパターンを見つけるための手法としてストリームマイニング技術が盛んに研究されている[1].

例えば，トレンドと呼ばれるデータストリームの特徴を分析する研究が存在する．トレンドは様々な形で活用され，過去から現在までのトレンドから未来のデータストリームの周期性を求め，データを予測することが可能である[2].

多くのストリーム予測手法では，予測対象のストリームの過去の計測値の傾向を用いて予測を行う．これらの手法ではデータストリームが過去の傾向と異なる変化をした場合は予測精度が悪化してしまう．しかし，室内に複数のセンサを設置した場合など限定された状況下では，センサからもたらされる複数のデータストリームの間に相関が発生しているケースが存在する．この場合，単独のデータストリームとしては過去の傾向とは異なる変化を示しているとしても，他のデータストリームとの間の相関は維持されているならば，相関を考慮することで予測精度の向上を図ることができると考えられる．

例えば，図 1のようにStreamAとStreamBが同時に計測されている状況を想定する．StreamBの計測値がある時刻から過去の傾向と異なる変化を示す場合, 過去の傾向のみを利用する予測手法では予測精度が低下すると考えられる．しかし，StreamAが変化した後に, StreamBがそれに追従する形で変化するといった相関があらかじめ分かっている場合，StreamBの予測を行う際にStreamAの変化を参考にすることで予測精度を向上させることができる．

本研究ではデータストリーム間の相関には次の 2 種類が存在すると考えている．

Ⅰ．Similarity Correlation

図 2のようなデータストリームAとBが計測されている場合を考える，このとき，ストリーム同士の計測値の差が小さいことから変化の傾向と計測値の類似性が高いと考えられ，ストリームの間には類似性に基づく相関があると考えられる．このような相関を本稿ではSimilarity Correlationと呼ぶ．

† 公立はこだて未来大学大学院
Graduate School of Future University Hakodate
‡ 公立はこだて未来大学
Future University Hakodate

ⓒ 2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.25
Vol.2012-UBI-33 No.25
2012/3/13

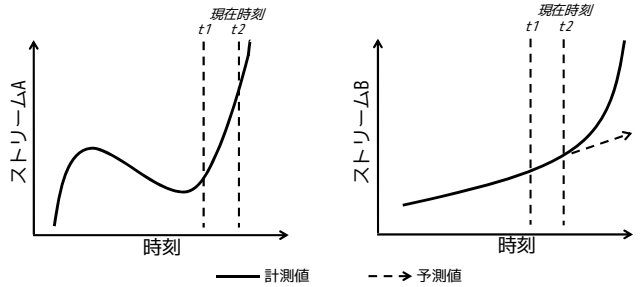

図 1 データストリーム間の相関の例

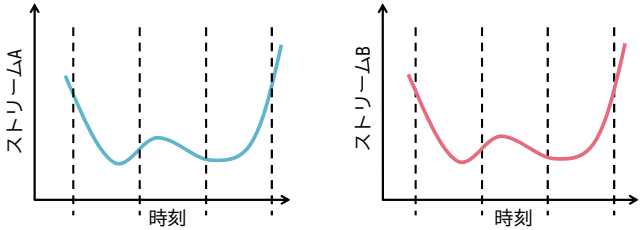

図 2 Similarity Correlation

Ⅱ. Timing Correlation

図 3のようなデータストリームCとDが計測されている場合を考える．この 2 本のデータストリームAとBの計測値の差は大きく，上昇や下降といった波形の変化も異なっている．しかし，データストリームAが上昇するとデータストリームBは下降するといったような，変化のタイミングの類似性が存在すると考えられる．本稿ではこれをデータストリームの相関の一種と考え，Timing Correlationと呼ぶ．

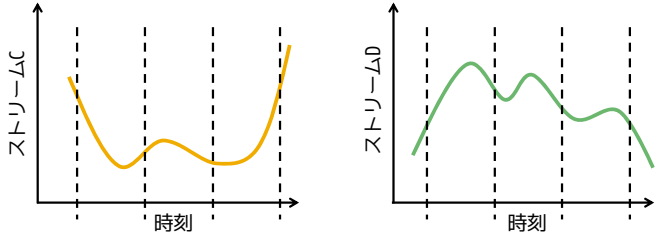

図 3 Timing Correlation

従来のデータストリーム間の相関検出手法では後者の相関を扱っているものは存在しない．しかし，Timing Correlation を用いることにより提案手法を適用可能なデータストリーム計測環境の範囲の拡大を図ることができると考えられる．相関の検出のため，提案手法はデータストリームを一定長の部分シーケンスに分割し，部分シーケンス毎にトレンド検出を行う．この部分シーケンスのトレンドはデータストリーム毎に大量に出現する．提案手法ではこのトレンド群をトレンド同士の類似性に基づいてクラスタと呼ばれるグループに分類する．これによりデータストリームの計測結果をクラスタ分類結果の流れによって表現することを可能にする．本稿では，このクラスタ分類結果を同時観測クラスタと呼称し，各データストリームのクラスタ分類結果の集合を同時観測クラスタ群と呼ぶ．同時観測クラスタ群は計測値ではなく，あくまでも計測値の変化の傾向を示す情報であることから，Timing Correlation のような計測値の類似性に基づかない相関の検出のために有用であると考えられる．

本稿は複数のデータストリームが同時に計測されており，かつ相関が存在する環境下においてストリーム予測精度を向上させることを目標とする．そのためにデータストリーム間の相関に基づいたストリーム予測を行うことを課題とし，同時観測クラスタ群とそれに基づく提案手法によりこれを解決する．

2. 関連研究

本章では，本研究と関連性の深いデータマイニングおよびストリームマイニングの分野における既存研究とその問題点について述べる．

2.1 トレンド検出

時系列データを特徴に基づいて分類した情報をトレンドと呼ぶ．データストリームも含めた時系列データから効率的にトレンドを検出することを目的とした研究が多く取り組まれている．川島ら[3]は APCA（Adaptive Piecewise Constant Approximation）と呼ばれる特徴量抽出手法を用いることで効率的な部分シーケンスの次元削減を行なっている．これにより計算量を削減することでデータベース内に存在するトレンドのサンプルデータとの高速なマッチング処理を行う手法を提案している．また，豊田ら[4]は，データストリーム毎のデータレートの違いや周期性の変化に対応することを目的としてダイナミックタイムワーピング（Dynamic Time Warping）距離に基づくトレンド検出処理を提案している．また，データストリームの部分的な類似を検出するために，データストリームの一部分を抽出した時系列データである部分シーケンスの類似判定手法を提案している．また，これの発展として消費メモリと計算時間を考慮した改良手法を提案している[5].

ⓒ 2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.25
Vol.2012-UBI-33 No.25
2012/3/13

2.2 相関検出

データストリーム間の相関を検出するための技術として櫻井らは BRAID[6]を提案している．BRAID は相互相関関数を用いて相関検出を行なう．相互相関関数の算出には平均や分散，内積が必要になるが，これらの統計情報はインクリメンタルに求めることが可能である．そのため，BRAID は新しいデータが計測される毎に相互相関関数を更新し，相関検出を行うことが可能となっている．また，Zhu らが提案した StatStream[7]はデータストリームの一部分を抽出した部分シーケンスに対して離散フーリエ変換(Discrete Fourier Transform)を適用することで DFT 係数を算出し，これを用いて相関値を求めることで部分シーケンス毎の相関検出を実現している．

これらの手法は Similarity Correlation については扱っているが，Timing Correlation については扱っていないことが問題点として挙げられる．

2.3 ストリーム予測

Papadimitriou[8]はトレンドにウェーブレット係数を用い，これを用いて自己回帰モデルに類似した線形モデルを作成する手法を提案している．ウェーブレット係数はデータストリームの周期性をよく表現するため，この手法は周期性の存在するデータストリームの未来の計測値を高精度に推定することができる．

ただし，周期性の存在しないデータ，あるいは周期性が変化するデータの推定には適さない．また，ストリーム間の相関についても特に言及されていない．

3. 要件定義

本稿はデータストリーム間の相関である Similarity Correlation と Timing Correlation に基づいたストリーム予測を行うことを課題とし，提案手法によってこれを解決する．まず，この課題に対して必要となる要件について述べる．

データストリームの予測は，過去のデータストリームの過去のトレンドから未来のトレンドを推定することで実現される．そこで，まずデータストリームからトレンドを検出する手法と，トレンド間の相関を抽出する手法が必要になる． Timing Correlation を検出するためには、DTW 距離や DFT 係数のようなデータストリームの計測値の類似度を用いる手法ではなく，データストリームの変化傾向の類似性に基づく柔軟な相関検出手法が必要とされる．さらに，その相関に基づいてトレンドの予測を行い，そこから計測値を推定する手法が求められる．

4. 提案手法

最初に，本提案方式が扱うデータストリーム環境について説明する．データストリームは複数同時に計測されていることが前提となる．これらのデータストリームにデータの欠損（データ計測・配送ミス）や遅延（データ計測・配送の遅れ）はなく，全て一定のデータレートによって継続的にデータが蓄積されているものとする．

データストリームの変化の傾向を示す特徴として，トレンドと呼ばれる情報が存在することはすでに示した．この考え方を利用して，データストリームのある短期間を抽出した部分シーケンスに対してトレンド検出を行うことで，データストリームを部分シーケンスのトレンドのシーケンスとしてみなすことができる．このとき，計測値が類似している部分シーケンス同士は類似したトレンドを持つと表現できることから，部分シーケンスはトレンドに基づいてクラスタリングすることが可能であると考えられる．つまり，部分シーケンスのトレンド検出結果をクラスタリング結果によって代替することができると考えられる．

本稿では，あるタイミングにおける計測されている全データストリームのトレンド検出結果をまとめて同時観測クラスタ群と呼称し，この同時観測クラスタ群のシーケンスをクラスタ遷移パターンと呼ぶ．本提案手法では，このクラスタ遷移パターンからデータストリーム間の相関を検出する．具体的には，あるデータストリームのペアの間で，あるトレンド検出結果が同時に得られるケースが複数回発生する場合，それらのトレンドの間に相関が存在すると判断し，これをデータストリーム間の相関とみなす．データストリームAとBの間に相関が存在する場合の例を図 4に示す．データストリームAのトレンド検出結果として与えられるクラスタが，A1，A2，A3，...，データストリームBの結果がB1，B2，B3，...となっているとする．これらのクラスタが頻繁に出現し，その出現タイミングがデータストリーム同士で一致している場合，データストリームAとBは相関していると判断する．図 4の場合，A1-A2-A3 の流れと B3-B2-B5 の流れがそれぞれ同じタイミングで頻出していることから相関が判断できる．

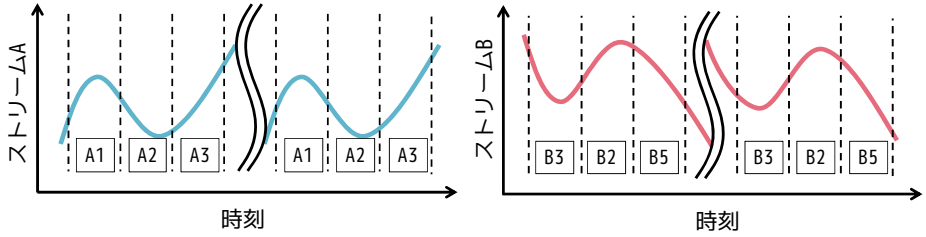

図 4 トレンドに基づく相関検出

クラスタ遷移パターンは図 4のそれぞれのクラスタの流れを内包する．本提案手法は，上述のようなクラスタ遷移パターンの出現回数を収集・蓄積することで全データ

ⓒ 2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.25
Vol.2012-UBI-33 No.25
2012/3/13

ストリームの相関情報の把握を行う．そのためのデータ構造としてクラスタ遷移テーブルを定義する．この例を表 1に示す．下表はデータストリームA，Bによる同時観測クラスタ群の，3 遷移分のクラスタ遷移パターンを蓄積した例を表している．リアルタイムに計測されるデータストリームから検出されるトレンドとクラスタ遷移テーブルに蓄積されている過去の相関情報をマッチングすることで，データストリームからリアルタイムに相関を検出することが可能となる．

過去のデータストリームと類似した計測値が最新のデータストリームで発生している場合，未来の計測値も過去のそれと類似したものになる可能性が高く，相関も維持されると考えられる．そこで，本提案手法は過去の計測値に加工を施して予測結果を作成する．例えば，データストリームA，Bの最新のトレンド検出結果が図 5のようにA3-A6，B2-B1 となっており，表 1のようなクラスタ遷移テーブルが得られているとする．このとき，A3:B2-A6:B1-A1:B5 というクラスタ遷移パターンが得られている場合，最新のトレンド検出結果とA6:B1 までの遷移が一致することから，未来の同時観測クラスタ群はA1:B5 と予想できる．この同時観測クラスタ群の予測結果から未来のトレンドが得られる．本提案手法は，予測されたトレンドから，そのトレンドに含まれると推測される計測値を抽出し予測結果としてアプリケーションに与える．

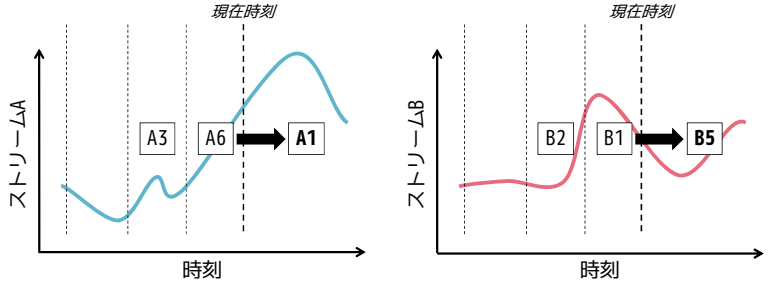

図 5 最新のトレンド検出結果とトレンド予測

表 1 クラスタ遷移テーブル

1		2		3		出現回数
A3	B1	A3	B3	B2	B4	4
A1	B4	A1	B2	A4	B6	6
A1	B3	A5	B5	A2	B3	12
A3	**B2**	**A6**	**B1**	**A1**	**B5**	**7**
…	…	…	…	…	…	…

最新のデータストリームのトレンド検出結果に基づいた相関検出を行うためには過去のデータストリームの計測値に基づいた相関情報（クラスタ遷移テーブル）を用意する必要がある．また，トレンド間の相関を正確に検出するためには，相関情報の作成に用いたトレンド検出結果と同様の基準で，最新のデータストリームのトレンド検出を行う必要がある．このためにはトレンド検出のための機構を作成・保存するための処理が求められる．このように，本提案手法はストリーム予測のために多くの事前処理を必要とする．そこで，本提案手法を相関検出のための事前処理と，それに基づいたストリーム予測処理の 2 つのプロセスに分け，次節以降で説明する．

4.1 相関検出のための事前処理

相関検出のための事前処理として，データストリームの過去の計測値の分割結果である部分シーケンスからトレンドを検出する機能が必要とされる．

前述のとおり，本提案手法では部分シーケンスの類似性に基づいてトレンド検出を行う．このとき，リアルタイムのトレンド検出と，過去の部分シーケンスに対するトレンド検出の双方が同様の基準で行われる必要がある．そこで，本稿は分類器を用いたトレンド検出手法を提案する．

分類器のための教師データとして，過去の計測値の部分シーケンスの分類例，つまりトレンド検出結果の例を与えることで，トレンド検出のための分類器を得ることができる．しかし，一般的なセンサネットワークなどから計測されるデータストリームを分類するための閾値や分類パターンがあらかじめ提供されている例は少ない．そこで，本稿では過去のデータストリームの部分シーケンスに対してクラスタリングを行い，部分シーケンス毎に分類例としてクラスタ番号を付与する．一連の流れを以下に示す．

i-1. 過去のデータストリームを部分シーケンスに分割
i-2. 部分シーケンスから特徴量を抽出
i-3. 特徴量をクラスタリング
i-4. クラスタリング結果をトレンド分類結果の例として教師データを作成
i-5. 教師データに基づいて分類器作成
i-6. 以上の処理を全データストリームに対して実行
i-7. 各データストリームの教師データを用いて，クラスタ遷移テーブルを作成

4.1.1 部分シーケンス分割と特徴量抽出

まず，データストリームを一定の時間軸で分割する．この一定の長さを持つ区間を窓（window）と呼び，分割されたデータストリームの一部分を部分シーケンスと呼ぶ．本研究では同じ窓で分割された部分シーケンスはそれぞれ同じ数のデータを持つものとする．また，同時に計測されたすべてのデータストリームは同じ数の部分シーケンスに分割される．

次に，部分シーケンスの波形的な特徴を効率的に表現する特徴量を抽出する．クラ

ⓒ 2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.25
Vol.2012-UBI-33 No.25
2012/3/13

スタリングは要素間の類似度に基づいてグループ分け（クラスタ分け）を行う手法であるが，その類似度の基準に部分シーケンスの特徴情報を与えることで，クラスタリング結果をトレンド検出結果とみなすことができる．本提案手法では複数の特徴量を併用することで部分シーケンスの分類性能の向上を図る．この特徴量の集合を特徴量パターンと呼び，これを各部分シーケンスから抽出する．

4.1.2 クラスタリングと分類器作成

特徴量パターンを用いて部分シーケンスのクラスタリングを行う．クラスタリング結果の各クラスタには特徴量パターンの抽出元の部分シーケンスが紐付けられ，各クラスタには類似した部分シーケンスが集められる．各クラスタにはクラスタ内に含まれる全ての部分シーケンスの平均値で構成されるセントロイドと呼ばれる部分シーケンスを設置し，これをクラスタの代表値として使用する．あるクラスタ内の部分シーケンスの様子を図 6に示す．このクラスタリング結果を元の部分シーケンスおよび特徴量と結びつけることで教師データを作成し，これを用いて分類器を作成する．

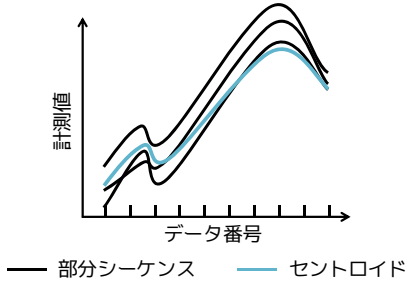

図 6 クラスタの内容

著書らの過去の研究から, k-meansなどの非階層型クラスタリングアルゴリズムを利用する場合，クラスタリング結果の中に，要素の過半数を含むような巨大なクラスタが形成される場合があり，これが分類器によるトレンド検出の精度を悪化させる原因となることがわかっている[9]．また，本提案手法はクラスタリング結果を教師データとして分類器を作成するが，分類器の精度を指標としてクラスタリング結果を調整する必要があることがわかっている[10]．そこで，本稿では階層型クラスタリングを用いて最適なクラスタリング結果を作成する手法を提案する．階層型クラスタリングはデンドログラムと呼ばれる木構造を作成する．部分シーケンスのデンドログラムの例を図 7に示す．

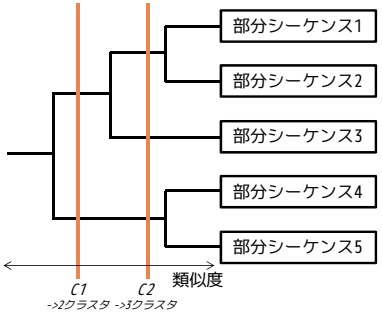

図 7 部分シーケンスのデンドログラム

図 7において，各部分シーケンスをつなぐ枝の長さが部分シーケンス間の類似度を表している．各枝の交差点はクラスタを表しており，この交差点間の長さがクラスタ間の類似度を表している．このデンドログラムに対して類似度の基準を与えることで，クラスタリング結果を抽出することが可能となる．例えば図 7において，*C1* でクラスタリング結果を抽出した場合, クラスタは 2 個となり, C2 を用いた場合は 3 個となる．この類似度を変化させることにより，一つのデンドログラムから様々なクラスタリング結果を抽出することが可能となる．このクラスタリング結果に対して，分類器の交差検定精度や各クラスタの要素数などの評価指標として与えることにより，分類器作成のための教師データとして最適なクラスタリング結果の選択を行う．

4.1.3 クラスタ遷移テーブルの作成

前述のとおり，本提案手法は過去のデータストリームにおけるトレンド間の相関情報の蓄積のために表 1のようなクラスタ遷移テーブルを用いる．クラスタ遷移テーブルは過去のデータストリームにおけるクラスタ遷移パターンを，一定の遷移数毎にまとめたものである．本提案手法は，遷移数の異なる複数のクラスタ遷移テーブルを作成する．これは遷移数の多い遷移パターンは，最新のデータストリームのトレンド検出結果と一致すれば相関が存在する可能性が高いのに対して一致する可能性が低く，逆に遷移数の低いものは相関を表現している可能性が低いが一致する可能性が高いためである．遷移数の多いクラスタ相関テーブルを優先的に検索することで，より強い相関を検出することが可能となる．

4.2 相関に基づいたストリーム予測

まず，ストリーム予測の全体のプロセスを以下に示す．ここで，*n* はある自然数とする．

ⓒ 2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.25
Vol.2012-UBI-33 No.25
2012/3/13

ii-1. 最新のデータストリームの同時観測クラスタ群の *n*-1 遷移分を用いて，クラスタ遷移テーブル検索のキーとなる遷移パターンキーを作成
ii-2. *n*-1 遷移分の遷移パターンキーを用いて，*n* 遷移分のクラスタ遷移テーブルから最も一致率の高い遷移パターンを検索
ii-3. *n* を変化させることで遷移数毎に分かれている各クラスタ遷移テーブルに対して ii-1，ii-2 の処理を行い，最も一致率の高いクラスタ遷移パターンを検索
ii-4. クラスタ遷移パターンを用いて未来のトレンドを予測
ii-5. 予測されたトレンドから未来の計測値を予測

クラスタ遷移テーブルの検索キーとなる遷移パターンキーは，クラスタ遷移テーブルの遷移数より 1 つ分短いものとする．例えば 3 遷移分の遷移パターンを持つクラスタ遷移テーブルを検索する場合，遷移パターンキーは 2 遷移分となる．この場合，一致率の判定は 2 遷移分までのパターンを用いて行われる．これによって，3 遷移目を未来のトレンドの予測結果として用いることが可能となる．この予測処理の様子を図 8に示す．図 8は表 1のような 3 遷移分のクラスタ遷移パターンを蓄積しているクラスタ遷移テーブルを検索する場合の例である．この場合の遷移パターンキーは A1:B3→A4:B2 であるが，クラスタ遷移テーブルの検索により，最も一致率が高い遷移パターンが例えばA1:B3→A4:B2→A3:B8 である場合，トレンド予測結果としてA3, B8 が導かれる．このトレンド予測結果から未来の計測値の予測が行われる．

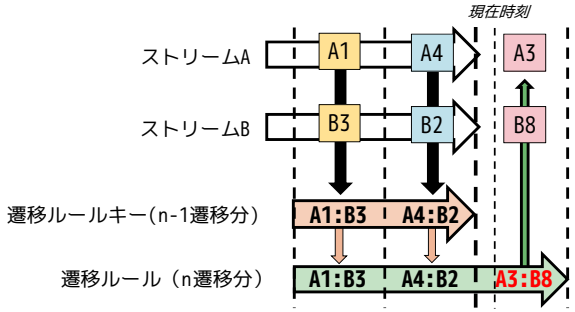

図 8 トレンド予測（n=3）

4.2.1 計測値の予測

トレンド予測結果を用いて，未来の計測値の予測を行う．各クラスタには関連付けられている全部分シーケンスを代表するセントロイドが存在する．セントロイドは全部分シーケンスの平均値であり，各部分シーケンスの特徴を表現した特別な部分シーケンスである．本提案手法では，クラスタから得られる予測値としてセントロイドを用い，このセントロイドの初期値からの変位と，最新の部分シーケンスの最新の計測値を結び付けることで予測値を算出する．例えば，温度センサの計測値のデータストリームであるデータストリームAの最新の計測値が 20℃であり，トレンド予測結果のセントロイドが図 9のような値を持っていた場合，20℃を初期値としてセントロイドに基づいた変位を持った部分シーケンスが予測値として与えられる．

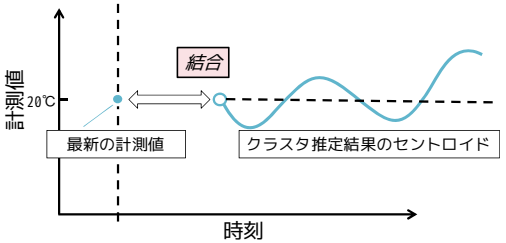

図 9 計測値の予測

5. 実験・考察

本提案手法は，過去のデータストリームの計測値に基づいて分類器とクラスタ遷移テーブルを作成し相関検出を行うことから，過去と同様の相関ならば，未来のデータストリームにおいても検出することが可能である．これは，データストリーム間の相関が維持されている限りは高い予測精度を維持できる可能性が高いということを意味する．これは，それぞれのデータストリームで見ると不規則な変化をしているとしても，計測されているデータストリーム全体としては過去と同様の相関が保たれているならば，提案手法の予測精度は悪化しないと考えられるためである．そこで，計測値の傾向が変化するようなデータストリーム群に対して本提案手法を適用した場合のストリーム予測精度について検証を行った．

5.1 実験環境

本提案手法を Java によって実装した．また，比較手法として R の自己回帰ライブラリを用いてデータストリームの過去の傾向のみに基づいたストリーム予測処理を実装した．また，今回は DBMS に既に蓄積されている時系列データをデータストリームと見なして実験を行った．これはリアルタイムに計測されるデータストリームを既存の DBMS などで取り扱うことは困難であり，実装が煩雑化するためである．

次に，実験に使用するデータセットについて述べる． まず，図 10のような部分シーケンスを作成する．それぞれ，左上がパターンA，右上がパターンB，左下がパター

ⓒ 2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.25
Vol.2012-UBI-33 No.25
2012/3/13

ンC，右下がパターンDである．それぞれ横軸がデータ件数，縦軸が計測値を表しており，それぞれのデータ件数は 24 件である．ここで，パターンAとパターンCの波形は類似している．また，パターンBとパターンDの波形も同様に類似している．これらの部分シーケンスを教師データ部と予測データ部でそれぞれ 44 本ずつ組み合わせることで， 2 本の時系列データの作成を行う．表 2の組み合わせはSimilarity Correlationを表現するよう，各時系列データの波形がなるべく類似するように組み合わせたものである．教師データ部ではある一定のパターン（A-A-A-BとC-C-C-D）が繰り返し出現するようになっており，予測データ部ではそれを無視した組み合わせが出現するようになっている．この 2 本の時系列データを合わせてSimilarityデータセットとする．次にTiming Correlationを表現するために，波形の類似性を無視するように表 3のような 2 本の時系列データの作成を行った．これら 2 本の時系列データを合わせてTimingデータセットとする．これらの 2 つのデータセットに対して，予測処理を行った．

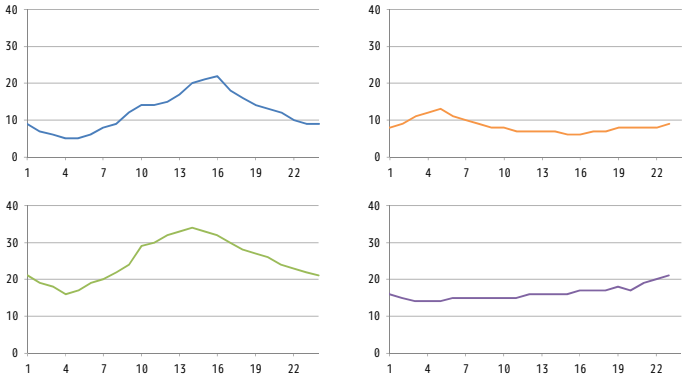

図 10 サンプルパターン A，B，C，D

表 2 Similarity データセット

	教師データ部								…	推定データ部								…
時系列データA	A	A	A	B	A	A	A	B	…	A	B	B	A	B	A	A	A	…
時系列データB	C	C	C	D	C	C	C	D	…	C	D	D	C	D	C	C	C	…

表 3 Timing データセット

	教師データ部								…	推定データ部								…
時系列データA	B	B	A	A	B	B	A	A	…	B	A	A	B	B	B	A	B	…
時系列データB	C	C	D	D	C	C	D	D	…	C	D	D	C	C	C	D	C	…

5.2 実験手順

提案手法による実験のプロセスを以下に示す．教師データを用いて分類器を作成

1-1. 分類器作成に用いたクラスタリング結果からクラスタ遷移テーブルを作成
1-2. クラスタ遷移テーブルに基づいて，予測データ部の部分シーケンス毎にトレンド予測処理を実行
1-3. トレンド予測結果から計測値予測結果を算出
1-4. 計測値予測結果と予測対象の部分シーケンスを比較し，差分値を算出
1-5. 全部分シーケンスの差分値の平均・分散を算出

以上のプロセスにおいて得られる部分シーケンスと計測値予測結果との差分は，実際に提案手法を用いて未来予測を行った場合の誤差にあたるものであり，その平均と分散は提案手法による計測値予測の誤差の期待値とみなすことができる．提案手法において部分シーケンスから抽出する特徴量として，平均値と分散，最大値，最小値と積分値を使用した．さらに周波数的特徴を抽出するために部分シーケンスの離散フーリエ変換結果の DFT 係数から上記と同様の各特徴量を抽出した．分類器作成アルゴリズムには，アンサンブル学習による高精度な分類器作成が可能である RandomForest を採用し，フリーのデータマイニングツールである Weka を利用した．

比較手法による実験手順を以下に示す．比較手法でも提案手法と同様に部分シーケンス毎に計測値予測を行うが，自己回帰モデルの精度向上のために，部分シーケンス毎の計測値推定を行う度に推定データから本来の計測値を抽出して教師データに結合し，自己回帰モデルの更新を行う処理を実装した．

2-1. 教師データを用いて自己回帰モデルを作成
2-2. 部分シーケンス毎に計測値推定を実行
2-3. 計測値推定結果と推定対象の部分シーケンスを比較し，差分値を算出
2-4. 全部分シーケンスの差分値の平均・分散を算出

また，テストデータセットの時系列データは 24 件分の部分シーケンスの集合であることから，計測値推定の際の部分シーケンス長はそれより短い 12 件分とする．

5.3 結果・考察

実験結果を表 4に示す．各データセットの予測誤差の平均において，提案手法のそれは比較手法を下回った．特にTimingデータセットにおいてその傾向が顕著である．実験前の想定通り，教師データと推定データで一貫した相関が存在する場合において，提案手法は比較手法を上回る高い推定精度を示すことがわかる．対して，比較手法にとっては表 2や表 3のような推定データはイレギュラーな波形でしかなく，自己回帰モデルによる推定に悪影響が出たために推定精度が低下したものと考えられる．

ⓒ 2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.25
Vol.2012-UBI-33 No.25
2012/3/13

表 4 ストリーム予測の精度

手法		提案手法				比較手法			
部分シーケンス長		12				12			
評価項目		平均	分散	最大値	最小値	平均	分散	最大値	最小値
Similarity	A	1.525	6.094	7.75	0	2.606	0.584	4.248	1.127
	B	2.036	10.295	9.583	0	2.944	1.997	6.645	1.379
Timing	A	1.358	5.809	7.75	0	2.444	0.425	3.502	1.565
	B	1.847	10.04	9.583	0	2.746	0.809	5.504	1.476

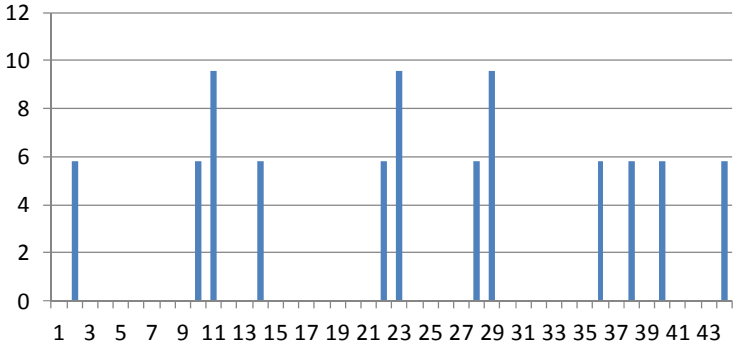

図 11 計測値推定誤差の変化

ただし，提案手法の予測誤差の分散は比較手法に対して非常に分散が大きい．このような大きな誤差は一定区間毎に出現した．Similarityデータセットにおける各部分シーケンスの精度の値を図 11に示す．上図において，横軸は部分シーケンスの番号であり，縦軸は誤差の値となる．このことから，提案手法はテストデータセットの元パターンの中のある区間の推定に失敗していることがわかる．これは提案手法が教師データ中のクラスタ遷移パターンに合わせたトレンド予測しかできないことが原因と考えられる．前述のように，教師データ部と推定データ部ではそれぞれ同時観測クラスタ群のクラスタ遷移パターンが異なるため，相関によって追従できない部分が出現してしまったものと考えられる．しかし，クラスタ遷移テーブルに遷移パターンが蓄積されているならばトレンド予測が可能となることから，トレンド推定が完了した部分シーケンスを教師データに追加してクラスタ遷移テーブルを更新することで精度悪化を抑制することが可能と考えられる．

6. まとめと今後の課題

本稿では，複数のデータストリームが同時に計測されている環境下において，データストリーム間の相関を用いることによりストリーム予測精度を向上させる手法を提案した．評価実験により，単一のデータストリームとしては過去の傾向にない動きをしながらも計測されているデータストリーム全体としては相関が維持されている計測環境下において，提案手法が高精度なストリーム予測を実現することを示した．今後はリアルタイムに計測されるデータストリーム計測環境において，実際に提案手法を適用し精度検証を行なっていく．

参考文献

[1] 有村博紀，喜田拓也，“データストリームのためのマイニング技術”，情報処理，Vol.46, No.1, pp.4-11 (2005).
[2] 藤原靖宏，櫻井保志，山室 雅司，“大量データストリームの類似探索手法”，日本データベース学会論文誌，Vol.4, No.4, pp.13-16 (2006).
[3] 川島英之，遠山元道，今井倫太，安西祐一郎，“波形特徴を用いた類似シーケンス検索”，情報処理学会研究報告．データベース・システム研究会報告，Vol.2002，No.67，pp.529-534 (2002).
[4] 豊田真智子，櫻井保志，市川俊一，“ダイナミックプログラミングに基づくストリームマッチング”，情報処理学会研究報告．データベース・システム研究会報告，Vol.2008，No.88，pp.277-282 (2008).
[5] 豊田真智子，櫻井保志，石川芳治，“部分シーケンスマッチングのためのストリームアルゴリズム”，電子情報通信学会論文誌.D 情報システム，Vol.J94-D, No.7, pp.1058-1070 (2011).
[6] S. Papadimitriou, A. Brockwell, and C. Faloutsos, “Adaptive, hands-off stream mining,” Proceedings of the 29th international conference on VLDB, pp.560-571 (2003).
[7] Y. Sakurai, S. Papadimitriou, and C. Faloutsos, “BRAID: stream mining through group lag correlations,” Proceedings of the 2005 ACM SIGMOD, pp.599-610 (2005).
[8] Y. Zhu and D. Shasha, “StatStream: statistical monitoring of thousands of data streams in real time,” Proceedings of the 28th international conference on VLDB, pp.358-369 (2002).
[9] 荒井健次，白石陽，高橋修，“クラスタ相関テーブルを利用したマルチデータストリーム予測方式”，電子情報通信学会技術研究報告．DE，データ工学，Vol.110，No.328，pp.55-60 (2010).
[10] K. Arai, Y. Shiraishi, O. Takahashi, “A Study on Stream Prediction based on Timing Correlation among Multiple Data Stream”, Proceedings of IWIN2011, pp.142-148 (2011).

ⓒ 2012 Information Processing Society of Japan