1R-08

# 多次元時系列データにおける オンラインでのモチーフ発見手法の検討

鷹取 留亞子† 上原 邦昭‡

神戸大学工学部情報知能工学科† 神戸大学大学院システム情報学研究科‡

## 1 はじめに

近年，時系列データから類似パターンを発見する研究が盛んに行われている．発見された類似パターンはモチーフと呼ばれ，時系列データの特徴を表す要素として重要な役割を持っている．モチーフの発見には，保存した時系列データをまとめて解析するバッチ処理 [1] と，データの生成に合わせて即座に解析するオンライン処理 [2] が存在する．オンライン処理には，使用する時系列の長さと計算時間の制約が存在する．一方，実世界において得られるデータは多次元となる場合が多い．より特徴を捉えたモチーフを発見するためには，多次元データの表現方法について考慮する必要がある．本研究では，多次元時系列データを対象として，オンライン処理によるモチーフ発見手法について検討し，実験を行う．

## 2 関連研究

時系列データの解析手段として，類似パターンの発見を用いる手法が多数報告されている．Lin ら [1] は，1 組のモチーフ集合をバッチ処理で発見する手法を提案している．一例を Figure 1 に示す．この研究では，類似度としてユークリッド距離を用い，計算の高速化や省メモリ化などについても実験を行っている．

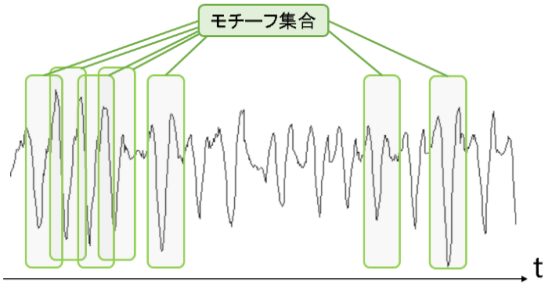


**Figure 1: バッチ処理によるモチーフ発見**

Mueen ら [2] は，オンライン処理に特化するために，探索窓を使用して，時系列データのうち最新の部分時系列から 1 組のモチーフ対を発見する手法を提案している．一例を Figure 2 に示す．

Online Discovery of Time Series Motifs in Multiple Time Series Data
Ruako Takatori, Department of Computer Science and Systems Engineering, Kobe University (†)
Kuniaki Uehara, Graduate School of System Informatics, Kobe University (‡)

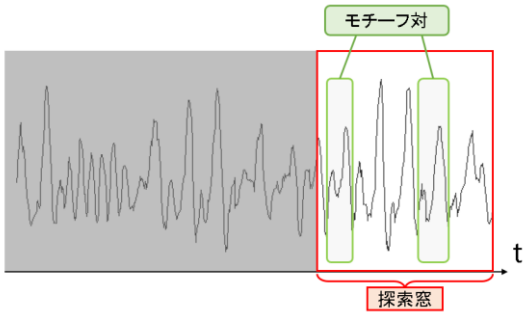


**Figure 2: オンライン処理によるモチーフ発見**

さらに Lam [3] らは，Mueen らの手法のデータ構造を改良し，$k$組のモチーフ対の発見を試みている．いずれの研究においても，類似度としてユークリッド距離を用いており，モチーフ長を様々に変化させ，計算時間や使用メモリ量に及ぼす影響について考察している．しかし，最適なモチーフ長の決定方法については言及されていない．

## 3 既存手法の課題

オンライン処理を行う既存手法の課題として，(1) 多次元データの扱いについて触れられていない点，(2) モチーフ長を実験者が人手で設定する必要がある点が挙げられる．

(1) 多次元データについて: Mueen らは，各次元でモチーフ発見を行った後，データの特性や実験目的に合わせた処理を行うべきとしている．しかし，人や動物の動きデータ等を扱う場合，対象の向きの変化を考慮すると，データの特徴が複数の次元にわたることが考えられる．Mueen らの手法では，複数の次元にわたるモチーフを捉えることはできない．

(2) モチーフ長の決定について: Mueen らや Lam らの研究では，最適なモチーフ長の決定を人手で行っている．したがって人の目で見つけられない潜在的なモチーフは発見できない．また，有用なモチーフ発見にモチーフ長が与える影響についても触れられていない．

これらの課題に対して，バッチ処理では様々な手法が提案されている．Tanaka ら [4] は (1) に対しては主成分分析 (PCA) を用い，(2) に対してはデータを SAX により符号化した後に最小記述長原理 (MDLP) を用いるアルゴリズムを提案している．

SAX は Lin ら [5] により提案されたデータの符号化手法である．SAX の例を Figure 3 に示す．

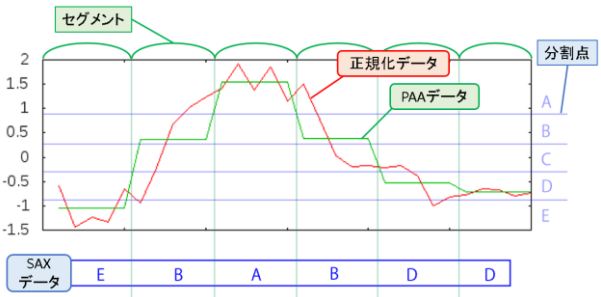


**Figure 3: SAX による符号化**

まず，実数値で与えられるデータを正規化し，時系列のセグメントに区切り，区間内の値の平均を取る処理を行う．これを PAA と呼ぶ．次に，正規確率分布において同じ大きさの領域となるように分割点を取り，PAA 後のデータが含まれる領域に応じて符号に変換する処理を行う．これを SAX と呼ぶ．Figure 3 では，6 個のセグメントに区切り平均を取った波形が PAA データとなる．さらに，4 個の分割点を取り，5 個の領域に符号 A, B, …, E を割り当てた結果，SAX データ”EBABDD”が得られる．SAX により，計算量削減とノイズ除去を行うことができる．

MDLP は Rissanen [6] により提案されたモデル最適化原理であり，データとモデルの記述長を表す MDL 評価関数が最小となるモデルを最適なモデルとする考え方である．また，得られた最適モデルはデータの特徴を最もよく表現していると仮定されている．Tanaka らは，符号化データ内に頻出するパターンを新たな一つの符号に変換するモデルの構築を提案している．

頻出するパターン$SC$の記述長$DL(SC)$は，$SC$の符号長を$n_p$，$SC$に含まれる符号の種類を$s_p$とすると，以下のように定義される．

$$DL(SC) = log_2 n_p + n_p log_2 s_p \quad (1)$$

また，符号列$\tilde{C}$に含まれるパターン$SC$を一つの符号に変換した場合，$\tilde{C}$の記述長$DL(\tilde{C}|SC)$は，変換後の$\tilde{C}$の符号長を$\acute{n_a}$，変換後の$\tilde{C}$に含まれる符号の種類を$s_a$，$SC$の出現回数を$q$とすると，以下のように定義される．

$$DL(\tilde{C}|SC) = log_2 \acute{n_a} + \acute{n_a} log_2(s_a + q) \quad (2)$$

これらの定義により，MDL 評価関数$MDL(\tilde{C}|SC)$は以下のように表される．

$$MDL(\tilde{C}|SC) = DL(\tilde{C}|SC) + DL(SC) \quad (3)$$

$MDL(\tilde{C}|SC)$を最小化するパターン$SC$をもとめると，最適モチーフ長は$DL(SC)$となり，モチーフとして$SC$が得られる．

## 4　提案手法

本研究では，オンライン処理により多次元データに対して PCA を行い，SAX と MDLP によってモチーフ長の決定とモチーフ発見を行うアルゴリズムについて検討する．また，1 組のモチーフ対の発見を対象とし，結果に基づいて$k$組のモチーフ対の発見手法についても検討する．

本研究に先立ち，バッチ処理によるモチーフ発見の予備実験を行った．予備実験では Mueen らの研究に基づいて，三次元の加速度データからバッチ処理によってモチーフ集合の発見を行った．また，PCA や SAX による符号化を行ったデータに対してもモチーフ集合の発見を行った．Figure 1 は，長さ 3,500 の三次元時系列データのうち，X 軸について符号化を行い，モチーフ長を 45 とした結果の一部である．

- モチーフ長を変えて複数回実験を行ったところ，モチーフ長によって得られるモチーフの数や最大類似度は大きく異なっていた．
- 各次元や PCA を行ったデータの結果を比較すると，時系列上でモチーフが多く発見される位置はそれぞれ異なっていた．
- 人の目ではモチーフと思われる部分でも抽出できていない場合が多くあった．
- SAX による符号化を行うと，正規化されて振幅の差が小さくなるため，行わない場合に比べて，発見されるモチーフの数が増加した．

以上を踏まえて，オンライン処理により多次元データに対して PCA を行い，SAX と MDLP によってモチーフ長の決定とモチーフ発見を行うアルゴリズムの実験と評価を行う予定である．

## 参考文献

[1] Lin, Jessica, Eamonn Keogh, Stefano Lonardi and Pranav Patel. "Finding motifs in time series." Proc. of the 2nd International Workshop on Temporal Data Mining (pp. 53-68). 2002.

[2] Mueen, Abdullah and Eamonn Keogh. "Online discovery and maintenance of time series motifs." Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1089-1098). 2010.

[3] Lam, Hoang Thanh, Toon Calders and Ninh Pham. "Online Discovery of Top-k Similar Motifs in Time Series Data." Proc. of 11th SIAM International Conference on Data Mining (pp. 1004-1015). 2011.

[4] Tanaka, Yoshiki, Kazuhisa Iwamoto and Kuniaki Uehara. "Discovery of time-series motif from multi-dimensional data based on MDL principle." Machine Learning, 58.2-3 (pp. 269-300). 2005.

[5] Lin, Jessica, Keogh Eamonn, Lonardi Stefano and Chiu Bill. "A symbolic representation of time series, with implications for streaming algorithms." Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (pp. 2-11). 2003.

[6] Rissanen, Jorma. "Stochastic complexity in statistical inquiry theory" World Scientific Publishing Co. 1989.