2B1-04

相関ルールとネットワーク分析による時系列データからの知識獲得

Knowledge acquisition from time series data by association rule and network analysis

金城敬太[*1†1] 尾崎知伸[*2†2] 澤井啓吾[*1] 古川康一[*2]
Keita Kinjo Tomonobu Ozaki Keigo Sawai Koichi Furukawa

[*1]慶應義塾大学環境情報学部
Faculty of Environmental Information, Keio University

[*2]慶應義塾大学大学院政策・メディア研究科
Graduate School of Media and Governance, Keio University

This research proposes a new method of time series data analysis which can contribute to elucidate the physical skill. This method extracts motifs (frequent patterns on time series data) by using the lift value used in association rule mining. As a result, we can extract motifs taking relations among other motifs into account. We extract association rules among acquired motifs, and then form a graph structure from those rules by using "the network analysis method" introduced by Yasuda. Finally, we demonstrate the feasibility of our method by analyzing respiration wave during cello performance with the method.

1. はじめに

近年，時系列データマイニング[Keogh 01]の研究が活発化している．そこでは，古典的な時系列データ解析手法のように時系列を直接使うのではなく，与えられた時系列を記号列に変換し、既存の記号処理手法を用いて特徴抽出を行う手法などが提案されている．また特徴的なパタンとしては，モチーフと呼ばれる時系列上の頻出パタン[Lin 02]などがあげられる．一方，社会学でも用いられているネットワーク分析法[安田 01]など，グラフマイニングを中心としたネットワーク研究も盛んに行われている．

本論文では，時系列上の特徴的なパタン発見問題に対し，モチーフを対象とした相関ルールの発見とネットワーク分析を併用するという，新たな手法を提案する．

以下に本論文の構成を示す．まず第二章で，モチーフを対象とした相関ルールの発見手法について説明する．ついで第三章で，得られた相関ルールへのネットワーク分析の適用方法について示す．第四章で実データを用いて提案手法の評価を行い，第五章でまとめと今後の課題を述べる．

2. モチーフとその逐次ルールの発見

本研究では，効率の良いモチーフ発見アルゴリズム EMMA (Enumeration of Motifs through Matrix Approximation) [Lin 02]をその出発点としている．時系列 $T = t_1, t_2, ..., t_n$ に対し，$T_m^l = t_m, t_{m+1}, ..., t_{m+l-1}$ を部分時系列と呼び，$T_m^l \in T$ と表記する．また，$d(T_a^l, T_b^l)$ を部分時系列 T_a^l と T_b^l 間の距離とする．EMMAアルゴリズムは，大雑把に言えば，時系列 T，長さ l 及び閾値 d を入力として，

$$|\{T_b \in T \mid d(T_a^l, T_b^l) \le d, (1 \le b \le n - l + 1)\}|$$

の値が大きいものから N 位までの $T_a^l \in T$ $(1 \le a \le n - l + 1)$ を高速に求めるアルゴリズムである．

この式からも分かるように，EMMAでは，頻出であるということのみをパタン抽出の基準としており，他のモチーフ(パタン)との関連は考慮されていない．また，モチーフ抽出の際に，予めモチーフ長や類似範囲(d)を与えなければならないという制約がある．これに対し本研究では，これらの制約を解消すると共に，より精密な特徴抽出を行うことを目的とし，相関ルールのリフト値を用いたモチーフ間逐次ルールの抽出と，そのネットワーク分析の併用を提案する．

連絡先:金城敬太 †1:現慶應義塾大学大学院政策・メディア研究科，〒252-8520，Tel.0466-49-3505，kinjo@sfc.keio.ac.jp
†2:現神戸大学 大学院自然科学研究科所属

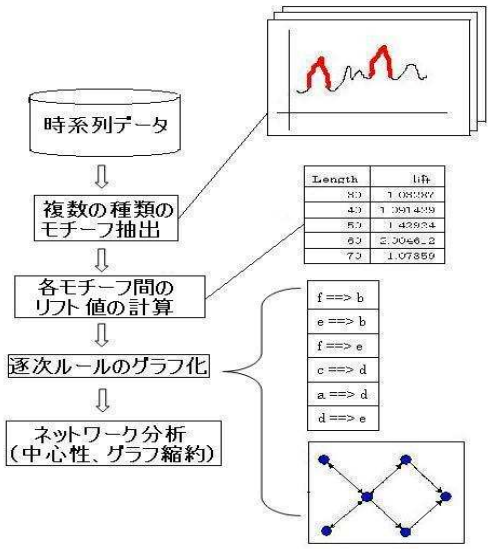


図1:本論文の手法

2.1 多種類のモチーフの抽出

提案手法の最初のステップは，最終的に得られるネットワークの構成要素の候補となるモチーフを発見することである．具体的には，以下の処理が行われる．

分析対象の時系列 T に対し，モチーフ長と類似範囲の組の集合 $P = \{\langle l_1, d_1\rangle, \langle l_2, d_2\rangle, \ldots, \langle l_m, d_m\rangle\}$ を与える．そして，EMMA アルゴリズムを用いて，各パラメタ $\langle l_x, d_x\rangle \in P$ 毎に T からモチーフの集合 M_x を求める．

2.2 リフト値によるモチーフの選別

提案手法の次のステップでは，相関ルール(逐次ルール)におけるリフト値を用いて，得られた各モチーフ集合 M_x $(1 \le x \le m)$ の順位付けを行う．

相関ルールを用いるためには，まず時系列データをトランザクション集合に変換する必要がある．モチーフの集合を M ，ウィンドウ幅を W としたとき，時系列 T のトランザクション集合 T_M^w を以下のように定義する．

$$T_M^w = \bigcup_{1 \le a \le n-w+1} \{\{ m \in M \mid d(m, T_b^l) \le d, (a \le b \le a+w-l) \}\}$$

提案手法では，この T_M^w から逐次ルールを導く．逐次ルールとは，ルールの前件と後件との時間的な関係を考慮した相関ルールである．具体的には，2 つのモチーフ T_a, T_b から構成されるルール $T_a \rightarrow T_b$ に対し，T_a の後に T_b が現れるという制約が課されている．また，T_M^w から得られる逐次ルールの集合を R_M^w で表す．

次に，リフト値[岡田 02]について簡単に説明する．リフト値とは，相関ルールの興味深さを表す指標の一つで，ルールの後件の条件付き確率と条件なし確率の比として定義される．すなわち，モチーフ T_a, T_b をアイテムとする相関ルール $T_a \rightarrow T_b$ のリフト値は，

$$P(T_a \mid T_b) / P(T_b)$$

として計算される．

提案手法では，上位 N 個のリフト値の平均が最も高いルール R_M^w に対応するモチーフ集合 M_x を最終候補として決定する．ここで，リフト値の平均を用いているのは，相関ルールを導くのにモチーフが適切に貢献していることを判別するためである．なお最終的には逐次ルールを導いているが，相関ルールでのリフト値を用いて貢献度を図ることは近似であると考えられる．

例えば，l を長くとれば重要かもしれないが少ないルールしか導きだせず，または l を短くとった場合には無意味なルールが多く導かれてしまうことが予想される．こういった問題に対処するために，提案手法を用いている．

3. ネットワーク分析による中心モチーフおよび縮約グラフの抽出

導出されたモチーフとその逐次ルールは，有向グラフにおけるノード及びリンクと考えることができる．すなわち，例えば，ルール $T_a \rightarrow T_b$ は，T_a, T_b がノードに，$\rightarrow$ がリンクにそれぞれ対応している．また置き換えた有向グラフは非対称の隣接行列によって表現できる．こうして表現した隣接行列を S とする．

今日，こうしたグラフ構造を分析するツールとして社会学関係で発展したネットワーク分析[安田 01]があげられる．ネットワーク分析では，例えばネットワークの中心性，縮約したグラフといった情報を導き出すことができる．

本研究では，このネットワーク分析の手法を，時系列データへと応用する．これにより，より精密な形で，モチーフ間からの情報抽出ができると考えられる．

まず，ネットワークの中心性を扱う．中心性とは，ネットワークもしくはシステム全体の中で重要性が高いノードを表したものである．具体的にはあるノードへのリンクの数の最大値として定義されたり，他にも隣接行列の固有値の最大値，媒介性，情報量の多さを基にしても定義される．本研究では，単純にあるノードへのリンク数（次数）を数えていき、その最大値をもとに中心性を抽出した．

次にグラフの縮約を行うことで抽象度の高い構造を抽出した．縮約したグラフは多次元の時系列を扱う場合など，記号やルールが多く，グラフが複雑になった場合に有用である．具体的には，ネットワーク上の構造，つまり行列の成分が類似するノードをまとめた行列の部分行列(ブロック B とする)を導きだす CONCOR アルゴリズムを用いた[安田 01]．流れを簡単に示す．隣接行列 S の列同士の相関係数を導き，それをもとに相関行列 S_{as} を作成する．さらにその行列の相関係数を計算していく．このことを繰り返すと，相関の値が収束する．ノードを表す成分同士が相関があれば，隣接行列 S の行と列を入れ替えて置換行列を作り，0と1の分布に従い S をブロック B の集合に分ける．こうして分割されたブロック B が値を持つならば1，0のみの場合は0に置き換えることで，縮約されたグラフを得る．今回はノードを統合するさい，単純に構造が類似であるということだけで判断するのは信頼性に問題があると考えた．そこで実際のモチーフのデータ間のユークリッド距離を用いたクラスタリングを行い，モチーフ全体と比較した中での近さも考慮に入れて統合するかどうかを判断した．モチーフの構造だけではなく，属性も含めたということである．

4. 呼吸データを用いた評価

提案手法の有効性を確認するために，提案手法をチェロ演奏時における呼吸データの解析へと適用した．具体的には，Bruch の Kol Nidrei 演奏時の呼吸データであり，被験者 1 名に 6 回演奏を行ってもらい，その平均を用いている．

呼吸の波形はそれが上昇した場合は息を吐く動作であり，また下降した場合は息を吸う動作として解釈ができる．その一部は図2に示した．なお，筋電図やモーションキャプチャなどの時系列データから有益な情報を得るという研究は，身体知の言語化[古川 05]といった観点からも重要である．提案手法では，相関ルールを導出する際に，ウィンドウ幅 W を与える必要がある．今回の実験では，時系列全体 T に対して自己相関関数 $r(q)$ を用いて W の決定をした．自己相関関数 $r(q)$ とは，全データ数を n ，ある時刻 t のデータを $x(t)$ として， t より q 時間遅れたデータ $x(t+q)$ との相関を計算するものである．式は以下のようになる．

$$r(q) = \frac{1}{n} \sum_{t=0}^{n-q-1} x(t+q)x(t)$$

この値が高いもしくは低い場合は，相関があるということになり，時系列 T は時間 q の大まかな周期であるといえる．計算をした結果，三つ目のピークがデータ数としては 133 となったのでこれを波形のピークが拾えるようにウィンドウ幅 W とした．この自己相関係数を用いたウィンドウ幅の決定方法は，周期を含む範囲からの特徴抽出を目的としている．一方，ウィンドウ幅の設定は例えば楽譜などのように時系列以外のコンテクストとの対応を考慮して設定することも可能である．

sublength	Lift
20	1.085306
30	1.252396
40	1.476613
50	なし

表1：モチーフ長とリフト値の平均

実験では，類似範囲 d を1．0に固定し，部分系列長 l を変化させながら，リフト値の平均を求めた．結果を表 1 に示す．

この結果より，リフト値の平均が最も大きい l =40 をモチーフ集合の最終候補として選択した．

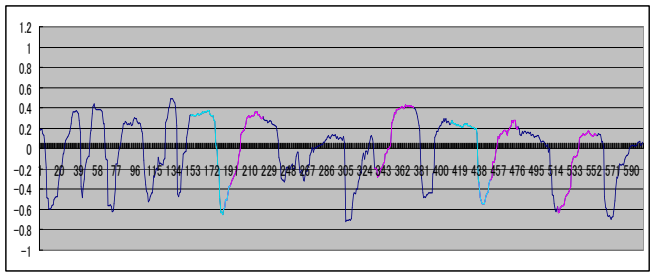


図2:呼吸波形とモチーフ

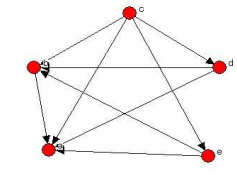
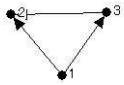


図3A:モチーフのネットワーク



B:縮約グラフ

結果として抽出されたモチーフ集合の一部分を，上記の図2に表示した.水色とピンクの部分がモチーフとなっている．呼吸の半分すなわち吐く，もしくは吸う動作がそれぞれ抽出されているということもわかった．さらに抽出されたルールに含まれるモチーフが，連なっている場合はひとつのモチーフとして捉えることもできる．例えば，図2において水色のモチーフの部分がピンクのモチーフの部分となっている箇所が 2 箇所ある．これにより，モチーフ長を固定したとしても，長さが大きいモチーフをルールとして部分的に抽出も行えるといえる．

図3Aはモチーフ同士の逐次ルールをグラフ化したものである．中心的なモチーフは，まず一番上のノードで出る次数が4つあることから出力される傾向が高く，左側の上下のノードは入る次数が3つであることから入力される傾向が高いことがわかる．また縮約したグラフを図3Bに示した．図3Aにおける右上下のノードが一つのノードになり，また左上下のノードが一つのノードとなり全体の抽象的な特長がつかめた．しかし，モチーフの数が少ないためあまり有用な情報が得られなかった．この手法はたとえば多次元にわたる関係などもっと大規模なネットワーク構造を扱う場合に有効であると思われる．

5. まとめ

本論文では，時系列データからのモチーフおよびそのルール抽出と，ネットワーク分析を適用する手法を提案するとともに，実データを使ってその有効性を示した．改善点としては，多次元時系列への適用や別のコンテクストとの関連を自動抽出して検証を行っていないことである．また，今回モチーフの長さを一定として，多様な長さのモチーフの関係を考慮に入れなかった．今後はそうした点も含めて検証していきたい．

参考文献

[Keogh 01] E..Keogh,: Mining and Indexing Time Series Data , Turorial at The 2001 IEEE International Conference on Data Mining(ICDM),2001.

[Lin 02] J. lin, E.keogh, P.Patel and S.Lonardi.: Finding Motifs in Time Series, In proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada. July 23-26,2002.

[岡田 02] 岡田孝:相関ルールとその周辺，第3回ORセミナー「データマイニングの実践と応用」，東京，特別講演 2002.

[古川 05] 古川康一:身体知解明へのアプローチ，第 19 回人工知能学会全国大会，2005.

[安田 01] 安田雪: 実践ネットワーク分析，新陽社 , 2001.