1H3-8

局所モデリング時系列データマイニングと帰納論理による知識獲得

Knowledge discovery by
local modeling time series data mining and inductive logic programming

金城敬太[*1] 澤井啓吾[*1] 古川康一[*1]
Keita Kinjo Keigo Sawai Koichi Furukawa

[*1] 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

Abstract This paper proposes a series of data mining algorithms to analyze time series data with non-stationary feature which are often observed in such data as EMG of human performance of skillful motion. Firstly we estimate a sequence of statistical models of a given time series datum, cluster them and allocate a symbol to each of them, Then, we extract relational patterns of the model expressed as a sequence of cluster symbols by using the inductive logic programming. The extracted patterns are expected to represent some important features of the given time series datum.

1. はじめに

近年、情報社会・ユビキタス到来により大量のデータが手に入るようになっている. それだけではなく、そのデータに伴う時間情報が手に入るようになり時系列データの解析が急務となっている. 特に従来のデータマイニングと組み合わせることで大量の時系列データを処理する時系列データマイニング手法は、重要性を増してくるだろう. そこで本研究は、時系列データマイニング手法の開発を主軸に、センサー情報から高次の情報処理を目指すシステムの開発を目的とした.

また、本研究では提案手法の具体的な応用分野として、スキルサイエンスをターゲットにした. これは人間の身体の暗黙的な知の言語化を目指している分野であり、センサーからの情報を取得して、人工知能・データマイニングの手法、認知科学の知見を用いて解明する新しい分野である. このスキルの言語化の過程と、時系列データマイニングにおける記号化という操作は親和性が高く、スキルサイエンスにおける高次の記号処理を通じたスキルの解明への橋渡しとなると考えられる.

2. 提案手法

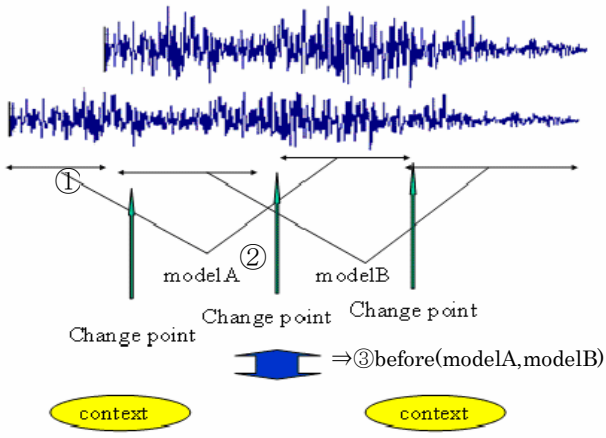
スキルサイエンスでは楽器演奏時における筋電などの分析において、一時系列内で複数のパタンがあり、さらにこれらが複数系列が存在する場合にどのようにルールを見つけるかという問題があった. このような問題に対しての時系列データマイニングの手法では線形近似や分節化などを行って分析をしているが、意味のあるパタンをどう抽出するかという問題があった. 一方、古くから局所的に変動する時系列データの分析に、局所定常自己回帰モデル[北川 05]やノンパラメトリック回帰における再帰分割線形回帰モデル[Hastie 02]、音声認識における線形モデルを使用した方法など数多くの手法が提案されている.

そこで本研究では、これらの知見を利用し、意味のあるパタンは似た時系列モデルで表現できると考え、それを抽出する一般的なアルゴリズムを考えた. また同時に本研究では時系列データマイニングにおける時系列データの分節化という問題にも情報量規準という視点から対処した.

また、これまでの研究で複数の系列でのモデル同士の関係を扱った研究はない. そこで研究では複数の非定常時系列における「時系列モデル間の関係」を抽出する方法の開発を行うことも研究目的とした. 時系列モデルを局所的に当てはめるだけではなく、クラスタリングし、クラスタごとに記号を割り振り記号処理に持ち込むことで、時系列モデル間の「関係」を扱った.

2.1 局所モデリング時系列データマイニング

提案手法について概要を述べ、次に具体的に説明する(図1). ①まず初めに時系列に対して情報量規準を使用して局所的に「統計的モデル」を推定し(再帰分割推定アルゴリズム)分節化、②それをクラスタリングを行う. ③次にクラスタごとに記号を割り当て、その関係ルールの抽出を帰納論理プログラミングを用いて行うという流れになっている.



提案手法

(1) 再帰分割推定アルゴリズム

再帰分割推定アルゴリズムとは、二分割アルゴリズムを再帰的に行う手法である. これにより、時系列データを時間的に異なる複数のモデルで表現出来る. 二分割アルゴリズムとは、一時系列データに対し分節点で切り、最適な二つのモデルで局所的に当てはめを自動的に行う方法である. 最適な分割点は、次の手法で得る. まず、ある時系列 T を分割点 $t(6 < t < length(T) - 6)$ で分割する. 次にその左右に対

連絡先:金城敬太, 慶應義塾大学大学院政策メディア研究科(現 総合研究大学院大学), 藤沢市遠藤 5322, Tel.0466-47-5350,kinjo@jmrlsi.co.jp

してあるパラメタ数内(今回は5)で時系列モデルを推定し、それぞれの最小の赤池情報量規準(AIC)を計算する. 最後に左右のモデルの最小 AIC の合計が最小になる時点を探すことで得られる. すなわち、時系列データを適切にモデルで表現している点で分割を行う. この分割後、さらに分割した左右に対して同様の分割を繰り返す. これが再帰分割推定アルゴリズムである. 停止の最低条件はモデルのパラメタ数がデータ数を上回るところになっている. 図1に簡略化したアルゴリズムを載せた. 二分割アルゴリズムは two_segment. 再帰アルゴリズムは regseg、各 models は推定したパラメタの値.

```
Algorithm  bestmodels=regseg(T)
  for i=6 to length(T)－6
   [two_sum_minaic  models]=two_segment(T,i)
  sumaics=[sumaics  two_sum_minaic]
   end
   minsumaic=min(sumaics)
   [segpoint best_models]=search(minsumaic,sumaics)
   if  minsumaic<one_model_aic
   best_models=regseg(T[1:segpoint])
   best_models=regseg(T[segpoint+1:length(T)])
   end
end
```

図１　再帰分割推定アルゴリズム

(2)　時系列モデルクラスタリング

次に一つの時系列データから得られた複数の時系列モデルをクラスタリングする. これにより、同じ系列で似た特長を持つ部分を抽出することが出来るのみならず、クラスタに記号を割り振ることで記号処理に持ち込むことが出来る.

まずクラスタリングをするために時系列モデル同士の距離を定義する. 今回、筋電の分析に用いられる自己回帰モデル(以下 AR モデル)を使用した. AR モデルは、筋肉の一活動状態に対応していると考えられる.

X が AR 過程に従う時系列データとすると以下のモデルで表現できる. π は係数. ε は誤差項, p は最大次数である.

$$X_t = \sum_{i=1}^{p} \pi_i X_{t-i} + \varepsilon_t$$

Picco[Picco 89]は、ARモデルに従う二つの系列の距離を、AR 係数のユークリッド距離を元に定義した. これ以外にも例えば周波数解析の結果を用いて距離を測る方法や確率モデルに従うのでカルバックライブラー情報量を用いた方法などが考えられるが今回は計算の簡便性を考えて、この距離を採用した.

次にクラスタリング手法について述べる. クラスタリングには数多くの手法が存在する. またクラスタ数を決める場合にも問題に依存して恣意的に決定する. しかし今回はクラスタ数について記号化・汎化を考慮し、自動的に決定する. こうした基準もいくつか存在しており、例えば Hastie らの GAP 統計量などが存在する[Hastie 04]. 本研究ではモデルの係数の距離をもとに k-means を使用し、Krzanowski,W.J & Lai,Y.T らが提案しているその群内平方和の平均の変化を表す KL 値が最大となる数でクラスタ数を決定した[Krzanowski 85]. W は群内の平均平方和の平均値である. すなわち、当てはまりの良さが大きく改善されているクラスタ数を採用している. これらは筋肉の状態数を表現していると考えられる. なお、正規化のために次元 p 分の2乗したクラスタ数をかけている(a).

$$DIFF(k) = (k-1)^{2/p} W_{k-1} - k^{2/p} W_k$$

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right| \qquad \text{(a)}$$

(3)帰納論理プログラミングによる時区間パタンの抽出

時区間パタンは、Allen の時区間論理[J.Allen 94]を用いる. 時区間論理とは、時間的な区間を持った対象同士の関係を記述するための体系である. 対象 x が y の前にある場合は before(X,Y), x が y と同時に終わるのであれば finished(X,Y) など7種ある. 例えば overlap(X,Y)は以下のようになる. Stime は対象の開始時点、Ftime は終了時点である.

```
overlaps(X,Y):-
    model(X,Stime1,Ftime1),model(Y,Stime2,Ftime2),
    Stime1<Stime2,Ftime1>Stime2,Ftime1<Ftime2.
```

ここで、時系列モデルは model(モデルの種類,モデルの開始時点,モデル終了時点)の順で記述をしている.

こうしたルールの集合、すべてを背景知識として述語論理の上で動く、学習器である帰納論理プログラミング[古川 01]を用いて頻出する時区間ルールの抽出を行った. これにより時区間論理で記述されたルールを導くことが出来る.

3. 実験

3.1 チェロ演奏からの情報抽出

チェロの熟達した被験者 1 名に対し、筋電前腕筋、上腕二頭筋、上腕三頭筋、三角筋前、三角筋後、脊柱起立筋-右、脊柱起立筋-左を計測し、課題として8、4、2、1拍子で D 線を演奏してもらった. ビートの違いによりどのような変化を起こすのかについて調べた. それぞれの試行回数は4とした. そこで各テンポを学習例(正の事例)としてルールを抽出した.

まずそれぞれの筋肉についての特徴の抽出のために局所モデリングを行い、時系列モデル群を抽出した. その後、モデルの係数を元にクラスタリングを行った. その際に、(1)の指標の変化を参考にクラスタ数を決定した. 例えば各筋肉のクラスタ数は以下表1のようになった. これらは筋電の活動のパタン数に対応する.

前腕筋	上二頭	上三頭	脊立筋-左	三 角筋前	三 角筋後	脊立筋-右
4	4	3	3	3	4	3

表1　各筋肉のクラスタ数

前腕筋についての詳細なデータをみると、それぞれのクラスタ内のモデル数は全部で 11 個で 3, 4, 1, 3 の 4 つのクラスタに分けられる. それぞれのクラスタに A,B,C,D を割り振った.

こうして得られた時系列モデルの記号をもとに、記号処理を行って時区間パタンの抽出を行う. まずすべての時系列モデルを述語論理の形式で記述する. 最終的に帰納論理プログラミングを実行し結果として得られたルールの例を示す. また χ 2乗検定(イエーツの補正)を行った.

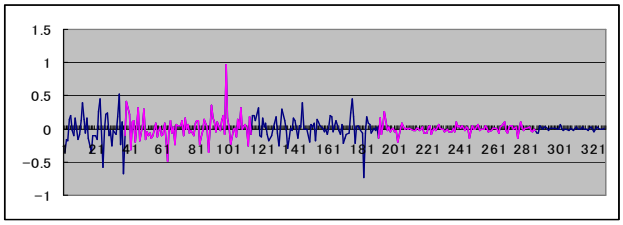
[8 拍子で得られたルール(Z=4.8132 $>x^2(1,0.05)$)]

```
beat(A) :-during(A, zenwan_A, jou_sn_B).
beat(A) :-overlaps(A, zenwan_A, jou_ni_C).
```

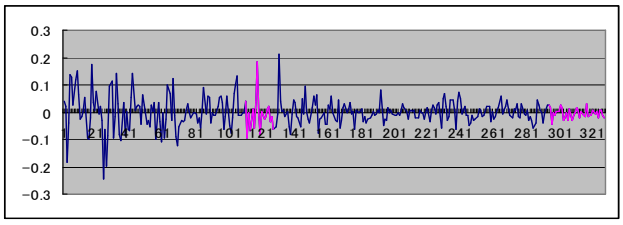
[1, 2拍子で得られたルール(Z=4.8132 $>x^2(1,0.05)$)]

```
beat(A):-before(A,zenwan_B,jou_ni_C)
```

これらを筋電の生データで見てみる．図2は beat(A):-before(A,zenwan_B,jou_ni_C)を示す事例である．それぞれ活動度が前後に比べて下がったときのパタンがルールとして検出されている．前腕の活動が下がった後に上腕ニ頭筋の活動が抑えられるということを示している．これは速いビートで演奏を行う場合は、腕の端から順に力を抜いて演奏を行っていることを示していると考えられる．それ以外の遅いビートにおけるルールでは、前腕の活性が長期に渡っており、その間で上腕三頭筋の活性が見られることが観測された．これらは、ビートを上げて演奏をする際に各筋肉のインピーダンスを調整して、振り子のような演奏をしている[古川 05]ことを示していると考えられる．



前腕筋



上腕ニ頭筋

図 2 筋電波形(青)と抽出されたパタン(赤)

3.2 バイオリン演奏の比較解析

次に、バイオリンの演奏の熟達者 1 名に対し、同様な実験を行い、バイオリンとチェロにおける分類ルールの学習を行うことでその違いについて調べた．

[バイオリン演奏全体のルール
(Z=13.9 $>x^2(1,0.05)$,posucover=5)]

```
beat (A) :- before(A, sekityuki2, sekityuki4).
beat(A) :-overlaps(A, soubo1, jowan_ni3).
```

バイオリン演奏におけるルールは、第一のルールは、脊柱起立筋の中での二つの活動パタンが検出されている．これは活動が抑えられている状態の後に、活動が活発になるという単純なパタンが検出されている．特に試行の最初の段階で検出がされている．第二のルールは、僧帽筋と上腕ニ頭筋での活動パタンである．それぞれ、活動が落ち着いているときの時系列データが検出されている．すなわち、僧帽筋の力を抜いた後に上腕ニ頭筋の力を抜くというパタンが検出されている．背中から徐々に力を抜いていることになる．

4. まとめ

局所的にパタンの変化する時系列データを分析する方法を提案し、さらに系列が複数ある場合などにその時系列モデル同士の関係パタンを抽出する方法を提案した．具体的にチェロ演奏およびバイオリン演奏の解析に適用した結果、意味の有るパタンの抽出もできた．

チェロ演奏においては速いビートで演奏を行う場合は、腕の端から順に力を抜いて演奏を行っていること、それ以外の遅いビートにおけるルールでは、前腕の活性が長期に渡っており、その間で上腕三頭筋の活性が見られることが観測された．これらは、事前に得られていた知識である振り子演奏とも対応関係がつく．また上腕ニ頭筋と上腕三頭筋の活動で活動が持続している場合があるが、これは腕を固定して引いているパタンを検出していると考えられ、特に全音符においてこの特徴がよく観察されることからもインピーダンスの調節を行って腕全体で引いている場面を観測していることが分かる．

バイオリンの解析においてはマクロなパタンとして、演奏が早くなればなるほど徐々に脊柱起立筋の活動が増えるということが観測された．また、チェロ演奏とは異なって、常に一定の活動を行う、すなわち筋肉・活動を固定した動きはあまり行わないといえよう．演奏する音符の違いを反映した演奏法の違いについては有用な知見が得られなかったが、細かい動きが多いため、セグメンテーションを詳細に行うために、サンプル数を増やすといった対処が考えられる．

本研究における今後の課題としては、被験者数が少なかったことが挙げられ、被験者を増やすことで個別人物の特徴的なパタンだけではなく、演奏についての一般的なルールを獲得することも出来る．また、セグメンテーションにおいて、ある状態が別の状態にすぐさま移るということを想定しているが、これについても時変係数を用いたり、局所的に周波数解析をかけることで対処していくことが出来ると考えられる．

参考文献

[Hastie 02] T.Hastie,Robert Tibshirani,J.Friedman The elements of statical learning ,springer 2002

[Allen 94]James F. Allen and George Ferguson, Actions and events in interval temporal logic. Journal of Logic and Computation, 4(5):531--579. 1994

[Krzanowski 85]Krzanowski,W.J & Lai,Y.T,A criterion for determining the number of groups in a data set using sum of squares clustering, biometirika 44, 23-34 , 1985

[Mark 04] Mark Last,Abraham Kandel &Horst Bunke data mining in time series databases, world scientific 2004

[Picco89]Picco.D,On a measure of dissimilarity betweenARIMA models.Proceedings of the A.S.A meetings-business and economic Stat. Washington D.C. 1989,1990

[伊庭 96]伊庭幸人“基礎的問題から見た情報統合”, 人工知能学会誌, 11 , 193-200,1996

[北川 05]北川源四郎『時系列解析入門』岩波書店 2005

[古川 01]古川康一、尾崎知伸、植野研(2001)『帰納論理プログラミング』共立出版 2001

[古川 05]古川康一、尾崎知伸、植野研 身体知解明へのアプローチ the 19th annual conference of the Japanese society for Artificial Intelligence 2005