The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ゆらぎ周期をもつデータのモデル化と決定木を用いた周波数解析

Modeling Stream Data with Fluctuated Periodicity and Periodicity Analysis Using Decision Tree

大場 孝二 Koji Ohba　木村 光樹 Mitsuki Kimura　渡邉 和哉 Kazuya Watanabe

株式会社本田技術研究所
Honda R&D Co., Ltd.

It is important to capture the periodicity of stream data in the field of natural sciences and real world applications. However, the periodicity often has not only some fluctuations but also noises. To find out the effect of such fluctuations and noises, we model the data and analyze. In data modeling, we apply the some randomness to express the fluctuations and noises. Then, we perform a Discrete Fourier Transform to the data and analyze these power spectrums using Decision Tree.

1 はじめに

気温の変化や潮汐や天気などの自然現象だけではなく心拍数といった生理現象等周期的な現象が数多く存在する[1]. 周期を特定することができれば予測やアノマリー検出を行うことができる. しかしながら, 現実世界において事象が完全に同じ周期で起こることはなく, 一定の幅で周期がゆらぐことが多い. さらに測定装置にノイズが混入することで本来あるべき周期が歪められることもある. したがって, ゆらぎやノイズを前提とした周期分析が求められる.

本稿では, 完全に周期的とはいえないゆらぎ周期をもつ事象を, ランダムデータ列を用いてモデル化した. さらにそのデータ列にノイズを加えた. こうして生成したデータ列に対して離散フーリエ変換を行い, パワースペクトルを得ることで, ゆらぎやノイズの影響がどのような形で表れるのかを明らかにした. さらにパワースペクトルを決定木によって分類することで, ノイズの影響がパワースペクトルのどの部分に影響を与えるのかを考察した.

2 データ生成と解析手法

2.1 データ定義

0 か 1 の二値をとる時系列デジタルデータを考える. a[x]を連続する x 個のデータ列で, 0 が x 個並んだデータ列とする. この時, 完全に周期的なデータ列を, ある整数 n を用いて [a[n-1], 1] が繰り返されているものとして定義する. 次にゆらぎ周期を持つデータ列を定義する. w[x]を連続する x 個のデータ列で, x 個のうち一か所のみが連続する x 個のデータの中で 1 で, その他は 0 となるデータ列とする. ただし, 1 となる箇所は完全にランダムである. この時, ゆらぎ周期を持つデータ列を, ある整数の組(q,p,m,n)を用いて, [a[q], [a[n-m], w[m]]*p]が繰り返されているものとして定義する. ただし, [a[n-m], w[m]]*p は[a[n-m], w[m]] を p 回繰り返したものである(図 1).

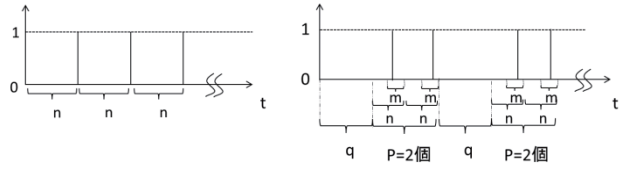

図 1 完全に周期的なデータ(左)とゆらぎ周期を持つデータで p=2 の場合(右)

連絡先:東京都港区赤坂 5 丁目 3 番地 1 号
Koji_Ohba@n.t.rd.honda.co.jp

2.2 生成するデータ列

生成するデータ列を以下のクラス 0,1,2 の 3 種類とする.

クラス 0 : 完全に周期的なデータ列, ゆらぎ周期をもつデータ列

クラス 1 : ランダムなデータ列

二項分布に従って 0,1 を各時刻にランダムに割り当てる. 各試行の成功確率を p とする.

クラス 2 : クラス 0 にクラス 1 を加えたデータ列

クラス 0 に属するデータ列に対して, クラス 1 に属するデータ列を加える. 値は各時刻のデジタル値を足して, 0 なら 0 のまま, 1 以上なら 1 とした.

2.3 分析手法

周期的な事象を判別するのに重要な特徴を分析するために以下を行った. まず, クラス 0,1,2 のデータ列を人工的に生成した. 生成したデータ列に対して, フーリエ変換を行い, パワースペクトルを得た. 得られたパワースペクトルデータを決定木にかけて分類し, 分類条件について分析した.

3 実験

本来であれば無限長のデータ列を解析しなければならないが, 今回は連続で 1460 個のデータが並んだデータ列を生成し, 分析した.

3.1 ノイズの影響によるパワースペクトルの変化

まず, クラスが 0,1,2 に属するデータ列を生成し, フーリエ変換を行い, 得られたパワースペクトルの可視化を行った. 生成したデータ列は, 以下の 5 種類である. クラス 0 のデータ列として, ①:n=28 の完全に周期的なデータ列と②:(q,p,m,n)=(24,1,4,4) のゆらぎ周期をもつデータ列の 2 種類を生成した. クラス 1 のデータ列として, ③:p=0.06 のランダムなデータ列を生成した. クラス 3 のデータ列として, ④:①に③を加えたものと⑤:②に③を加えたものを生成した. これら 5 種類のデータ列とそれらにフーリエ変換をして得られるパワースペクトル結果を図 2,3,4 に載せる.

図 2 上はデータ列①の, 図 2 下はデータ列②の結果である. 図 2 上のパワースペクトルは周期に相当する振動数(frequency) 1/28 の箇所にピークが立っている. 1/28 の整数倍にもピークが立っているが, これは 0,1 という離散データをフーリエ変換したためである. 一般に離散データに関しては, ある特定の振動数の整数倍の位置にピークが立つ. 図 2 の下図のパワースペクトルも同様に 1/28 の位置にピークが立っているが, スペクトル値の大きさが異なる. 1/28 の整数倍のピークが, 右にいくにつれて

小さくなっている．これは，ゆらぎ周期に導入したランダム列 w[x]の影響によるものである．小さな振動数帯では 28 という大きな周期を反映して 1/28 にピークが立つが，w[x]を反映して振動数が大きくなるとノイズのようなパワースペクトルとなる．

図 3 はデータ列③の結果である．パワースペクトルに特徴的なピークが見られず，特徴的な周期を持たないことがわかる．ランダムノイズでは，パワースペクトル値が全体に均等に散らばる．

図 4 上はデータ列④の，図 4 下はデータ列⑤の結果である．図 2 と比較すると同じ振動数のところにピークが立っているが，ピーク自体は小さくなっている．ランダムノイズがパワースペクトル値の値を全体的に均等にする役割を果たすからだと考えられる．

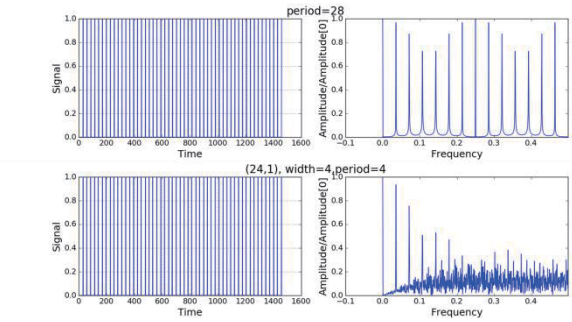

図 2 クラス 0 の 時系列データ（左）と パワースペクトル（右）．
上図はデータ列①，下図はデータ列②．

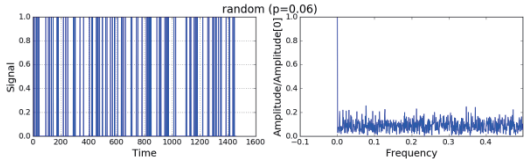

図 3 データ列③．

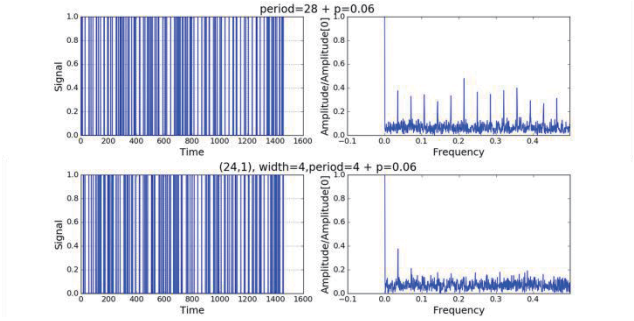

図 4　上図はデータ列④，下図はデータ列⑤．

3.2 決定木による分類結果

データを生成して，それぞれのデータのパワースペクトルを決定木にかけ，クラス分類に必要なパワースペクトル値を明らかにした．生成したデータは，クラス 0,1,2 それぞれ 1600 列である．クラス 0 は 2 種類の完全に周期的なデータ列と，3 種類のゆらぎ周期を持つデータで構成される．完全に周期的なデータは n=4, 28 で，ゆらぎ周期を持つデータは(q, p, m, n) = (8, 20, 1, 4), (20, 2, 4, 4), (24, 1, 4, 4)である．それぞれを 320 列ずつ生成した．クラス 1 は p=0.01, 0.02, …, 0.08 の 8 種類のランダムなデータ列をそれぞれ 200 列ずつ生成した．クラス 2 は，クラス 0 の 5 種類それぞれにクラス 1 の 8 種類をそれぞれ加えた 40 種類をそれぞれ 40 列ずつ生成した．以上のように生成したデータ列に対してフーリエ変換を行って得られたパワースペクトルを特徴として決定木によってデータ列のクラス分類を行った．分類の指標にはジニ係数を用いた．完全に分類できるまで決定木を動作させた結果を図 5 に載せる．

図中の各ノードについて説明する．Samples は分類対象となるデータ列数であり，value に各クラスのデータ列数が表示されており，左から順にクラス 0,1,2 のデータ列数である．gini はそのノードに含まれるデータ列のジニ係数である．ノードの分割条件は各ノードの一番上に書かれており，ルートノードでは振動数 (frequency) が 1/365.0 の時のパワースペクトル値が 0.001 以下であるかどうかが条件となっている．この条件を満たすデータ列は，図中で左下のノードに分類され，そうでないデータ列は右下のノードに分類される．

第一分岐でクラス 0 をほぼ分類できた．第一分岐では，振動数 1/365.0 でのパワースペクトル値の大小によってクラス 0 とクラス 1,2 を分類する．その振動数に注目したのは，クラス 0 は低振動数領域のパワースペクトル値が小さいが，ランダムノイズの影響によってクラス 1 のその領域のパワースペクトル値が大きなったことだと考えられる．第二分岐では，振動数 1/28.0 の振動数のパワースペクトル値の大小によってクラス 1 とクラス 2 を分類する．これは，クラス 1 が 28 という周期を持つためだと考えられる．第三分岐まで行うと，クラス 1,2,3 をほぼ完全に分類できた．

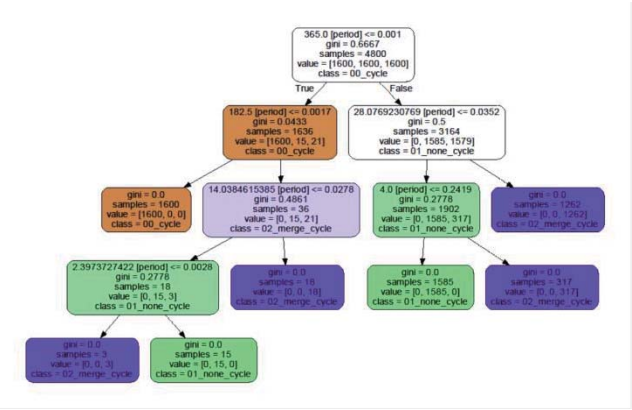

図 5: クラス 0,1,2 の決定木による分類

おわりに

本研究では，ゆらぎを持つ周期データ列をモデル化し，それらのデータ列に離散フーリエ変換を行い，結果を決定木で分類した．フーリエ変換後のパワースペクトル値はノイズの影響で減少する．パワースペクトルの値を見ることで，完全なノイズデータとノイズを加えた周期データの分類ができることを確認した．決定木の分類条件を確認することで，データ列がどのような周期を強く持っているのかを判定することができる．今後は，実データを用いた解析を行っていきたい．実際のデータには，ランダム性が成功確率 p の二項分布に従うものだけでなく，ほかの確率分布に従うものもあることが予想される．そのような状況においても，どの周期を強く持っているかの判定ができるかかの解析を行いたい．

参考文献

[1] VLACHOS, Michail; YU, Philip; CASTELLI, Vittorio. On periodicity detection and structural periodic similarity. In: *Proceedings of the 2005 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics, 2005. p. 449-460.