1F2-02

モチーフによる肝炎波形データからの特徴抽出の有効性に関する考察

Extracting Feature based on Motif from a Chronic Hepatitis Dataset

北口 真也[*1] Shinya Kitaguchi　佐藤 芳紀[*1] Yoshinori Sato　阿部 秀尚[*2] Hidenao Abe　大崎 美穂[*3] Miho Ohsaki　山口 高平[*4] Takahira Yamaguchi

[*1]静岡大学大学院情報学研究科 Graduate School of Informatics, Shizuoka University

[*2]静岡大学大学院理工学研究科 Graduate School of Science and Technology, Shizuoka University

[*3]同志社大学工学部 Fuculty of Engineering, Doshisha University

[*4]慶應義塾大学理工学部 Fuculty of Science and Technology, Keio University

At the stage of data pre-processing in knowledge discovery, feature extraction is so crucial. In this paper, given time-series data from a chronic hepatitis data set, we take Motif developed by Keogh in order to support a medical expert in discovering interesting knowledge. After improving Motif-based method with adding level values. Compared refined Motif-based method with ordinary K-means-based method, case studies show us that the former works well to do so.

1. はじめに

我々は，時系列データからの知識発見のために必要な前処理として，データ洗浄・検査項目の選択・周期の均一化・補完・離散化という処理を検討 [Ohsaki02] し，離散化の手法として，数値の波形データに直接 K-means を適応してきた．しかし，この離散化の手法では，各クラスタの代表とクラスタに含まれる生データとの乖離が，専門家が興味を持つ知識発見支援の妨げになるという問題がある．この問題は縦軸が数値，横軸が時間という時系列の軸上での波形間の距離でクラスタリングをしている弊害により起こると考えられる．一方，専門家は波形を評価するとき，興味のある特徴に注目することがある．すなわち，我々の目的のためには，特徴空間内でクラスタリングを実行する必要がある．

本稿では，波形変化に注目するモチーフを抽出する方法を紹介し，モチーフの問題点である波形のレベル値の問題を改善し，モチーフによるクラスタリングを提案する．さらに，慢性肝炎の病状の変化を示す重要な検査項目である GPT の反復性が各クラスタの代表に見られるかどうか，また，反復性を持つ GPT データがどのようなクラスタに分布しているのかという観点から比較評価する．

2. モチーフ抽出法とその問題点

本節では，波形パターンを抽出する方法としてカリフォルニア大学リバーサイド校の Keogh らのグループが開発したモチーフ抽出法 [Lin02] についてを紹介し，その後，問題点について述べる．

2.1 モチーフ抽出法

モチーフ抽出法は，入力時系列データ (波形) の一定期間の特徴を記号化し，特徴を示す記号値を比較することによりモチーフ (共通した特徴を持つ波形パターン) を抽出する手法である．モチーフを抽出するためには正規化，波形の分割，量子化，量子化値の比較という 4 つの処理を行う．まず，入力時系列データからサブシーケンスを切り出し，切り出されたサブシーケンスに対して正規化を行う．この正規化の処理は標準化であり，1 サブシーケンスに対してそのサブシーケンスの平均値が 0 となるように各サブシーケンスを構成する値を正規化する処理である．次に，正規化 (標準化) された波形をユーザーから指定された任意の数に分割する．各分割区間の値は，分割区間に含まれる値の平均値とする．そして，標準正規分布により求められる量子化の境界線によって，分割区間を示す値を量子化を行う．標準正規分布とは，平均を 0，分散を 1 とした正規分布である．なお，量子化数はユーザーが設定する．量子化後，各サブシーケンスは各分割区間に対応する記号列に置き換わる．比較の方法は，まず，各分割区間毎に比較し，量子化値の絶対値の差の総和を算出する．最終的に各サブシーケンスの関係がモチーフであるかどうかは，ユーザーが指定した閾値により判断する．

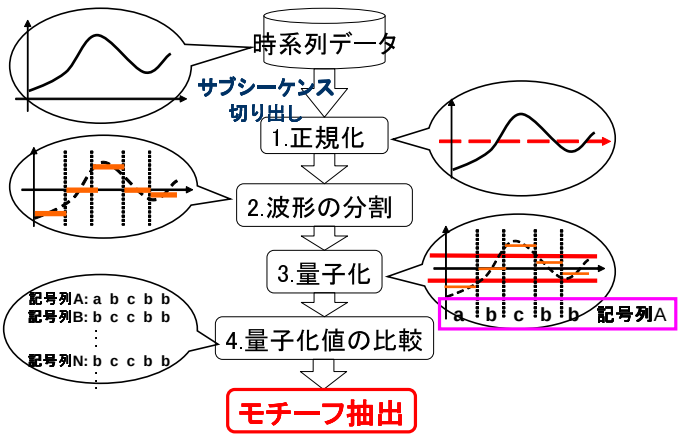

図 1: モチーフ抽出法

2.2 モチーフ抽出法の問題点

このモチーフ抽出法の問題点は，正規化により波形のレベル値の情報が失われることである．この問題により，波形のレベル値が異なる状態でもサブシーケンスの変化パターンが同じであれば，モチーフと判断される．しかしながら，本稿のケーススタディ (肝炎波形データからの知識発見支援) では，変化のレベル値の情報が重要であるため，そのレベル値の情報を付

連絡先: 北口 真也，静岡大学大学院情報学研究科，〒 432-8011 静岡県浜松市城北 3-5-1，Tel: 053-478-1473，FAX: 053-473-6421，e-mail: kitaguchi@ks.cs.inf.shizuoka.ac.jp

加する方法をモチーフ抽出法に追加する．付加情報は，全体のデータの平均値を基準として，各サブシーケンスの平均値を量子化した値である．この付加情報と分割区間の特徴を示す量子化値とは，全く意味が異なるため，重み付けによって区別して扱う．その重み付けはユーザーが与えるものとする．

これにより，モチーフのパラメータには，波形データにおける時間軸の分割数を設定する波形分割数，波形が持つ特徴を量子化値に変化させるときの量子化値の範囲の設定をする量子化値数，各サブシーケンスがどこまで類似していればモチーフとみなすかという閾値という 4 つと，変化のレベル値への重みを加えた 5 つのパラメータを設定することになる．現在はこの 5 つのパラメータを試行錯誤により設定している．

3. モチーフによるクラスタリング

2.1 節の手法では，モチーフの関係が 2 項関係で表せるが，多項関係でモチーフの関係を表すことが難しい．我々は，まず，各サブシーケンスのモチーフの関係を結びつけ，クラスタを作成できるだけ作成する．そして，作成されたクラスタ同士を合併することにより指定されたクラスタ数まで，クラスタ数を減少させる手法を以下に述べる．

3.1 モチーフからのクラスタ作成手法

モチーフからのクラスタ作成手法は，モチーフの 2 項関係を基にして，3 段論法のように同様の特徴を持つモチーフを同士を同じクラスタに含めていく手法である．図 2 にモチーフからのクラスタ作成手法の詳細を示す．各クラスタは木の構造を持ち，木のルートがそのクラスタの代表となる．木の生成では，親ノードと子ノードがモチーフの関係を保ちつつ，レベルを増やしていく．しかし，木のレベル数を指定しなければ，すべてのデータが 1 つの木で表現に含まれてしまい，意味がない．そこで，木のレベル数を指定する必要がある．ただし，各データは 1 つの木に含まれるようにする．

```
モチーフからのクラスタ作成
  For n= 1 to No_of_subsequence
      最も多く他のデータとモチーフの関係を持つ
      and どのクラスタにも属していないデータを抽出;
      抽出されたデータを新規クラスタの作成・代表に;
      For l = 1 to 指定レベル数
          クラスタに含まれたデータとモチーフの関係を持つ
          and どのクラスタにも属していないデータを同じクラスタへ;
      end_for
  end_for
  作成されたクラスタを出力
```

図 2: モチーフからのクラスタ作成のアルゴリズム

3.2 クラスタの合併手法

3.1 で作成したクラスタを図 3 に示す手順によりクラスタを合併する．この手法では，まず，クラスタに含まれるデータ数が最小であるクラスタを探索する．次に，クラスタ間の類似性をより，合併するクラスタを決定する．クラスタ間の類似性は，クラスタの代表とモチーフの関係であるデータを各クラスタがどれだけ含んでいるかによって調べることができる．すなわち，データ数が最小のクラスタはそのクラスタの代表とモチーフの関係であるデータが最も多く持つクラスタに吸収される．また，クラスタの代表とはそのクラスタ内のデータと最も多くのモチーフの関係を持つべきであるため，クラスタの合併後は再設定を行う．以上の処理をユーザーから指定されたクラスタ数に減少するまで行う．ただし，全クラスタへの探索が終了した場合は，この処理を打ち切る．

```
クラスタの合併手法
  total = first_total = 全クラスタ数;  n = 0;
  while total > 指定クラスタ数 && n <= first_total
      データ数が最小のクラスタ A を探索;
      クラスタ A の代表とモチーフの関係である
      データを最も多く含んでいるクラスタ B を探索;
      クラスタ A をクラスタ B に合併;
      クラスタ B の新代表の設定
        (クラスタ B 内で最もモチーフの関係を持つデータの探索);
      total = 全クラスタ数;  n + +;
  end_while
  作成されたクラスタを出力
```

図 3: クラスタの合併手法

4. 性能評価実験

本節では，実際に慢性肝炎データを用いて，従来のクラスタリング手法である K-means とモチーフによるクラスタリングを比較し，その有用性について検討する．

本実験における比較評価基準は，GPT の反復性が各クラスタの代表に見られるかどうか，また，反復性を持つ GPT データがどのようなクラスタに分布しているのかという観点から比較評価する．

4.1 実験概要

使用データは，千葉大学医学部附属病院から提供された慢性肝炎データである．その慢性肝炎データの中から慢性肝炎の病状把握に重要な検査項目であり，かつそのデータが持つ反復性に専門家が興味を示している GPT データを使用する．GPT データの反復性はデータを専門家や非専門家が実際に目で見て確認している (図 4) が，医学的根拠がまだ発見されていない．

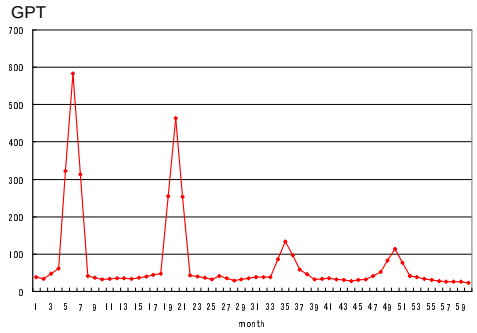

図 4: 反復性がある GPT データ

実験データは GPT の全患者データから切り出し期間 5 年，スライド幅 2.5 年でサブシーケンスを切り出し，欠損値が含まれていないデータを使用する．実験データ数は 724 であり，その中で反復性を持つデータは 21 データであった．次に，使用するクラスタリング結果の条件を設定する．まず，K-means では，各クラスタに含まれるデータ数が 10 データ以上であり，かつ各クラスタの代表と各クラスタに含まれるデータの乖離量が最小であるものを使用する．この条件による K-means でのクラスタ数 8 のクラスタリング結果は，図 5 である．図 5 は各クラスタの重心を示す．

モチーフによるクラスタリングでは，K-means で使用した 2 つの条件に加えて，クラスタ数 8，全クラスタに含まれるデータ数が全実験データの 8 割以上というパラメータを設定する．モチーフによるクラスタリング結果を図 6 に示す．図 6 は各

クラスタの代表となる各クラスタに含まれる実際の 1 データを示している．この時のモチーフによるクラスタリングのパラメータは，波形分割数：6，量子化数：5，モチーフとみなす類似度：2 以下，木のレベル数：2，波形のレベル値への重みは 3 であった．

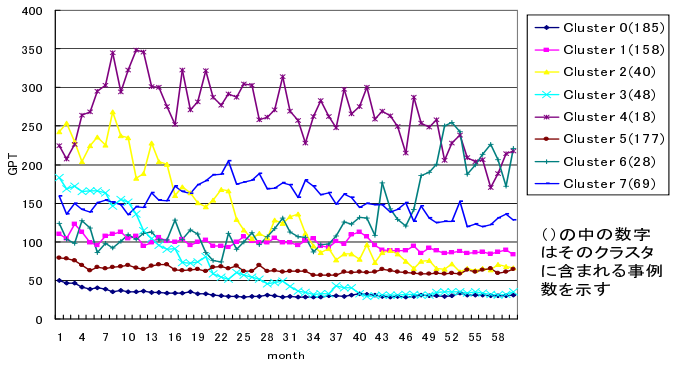

図 5: K-means によるクラスタリング結果

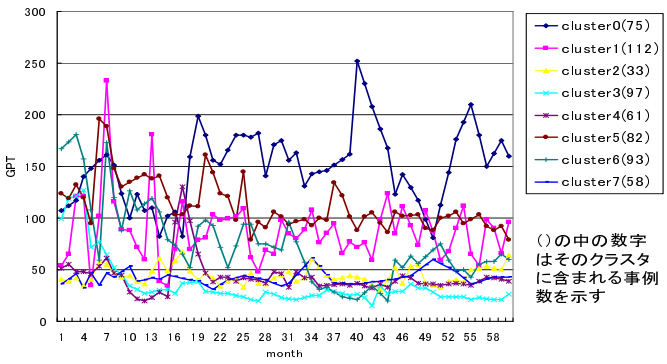

図 6: モチーフによるクラスタリング結果

また，GPT の反復性を示す 21 データのうち，K-means とモチーフによるクラスタリングから作成されたどのクラスタにいくつのデータが属するのかを表 1 に示す．

4.2 性能評価

まず，各クラスタの代表が反復性を示しているかという観点で評価する．図 5 より K-means では，クラスタの代表に反復性を見ることができるのはクラスタ 2，クラスタ 4 であるが，図 4 のようなピークが明らかな波形ではない．一方，図 5 よりモチーフによるクラスタリングでは，クラスタの代表に反復性を見ることができるのは，クラスタ 0，クラスタ 1，クラスタ 4 であり，K-means よりピークがより良く示されている．以上より，各クラスタの代表に現れる反復性の可読性においては，モチーフによるクラスタリングの方が有効であるといえる．

次に，反復性を示すクラスタにどの程度反復性を持つデータを含まれているかによって評価する．K-means では反復性を示すクラスタ 2，クラスタ 4 には，表 1 より，3＋1＝4 データを含む．一方，モチーフによるクラスタリングでは反復性を示すクラスタ 0，クラスタ 1，クラスタ 4 には，表 1 より，4＋2＋2＝8 データを含む．よって，モチーフによるクラスタリングの方が有効であるといえる．

表 1: 反復性を持つ GPT データの分布表

クラスタ番号	K-means	モチーフによるクラスタリング
クラスタ 0	2	4
クラスタ 1	4	2
クラスタ 2	3	1
クラスタ 3	0	2
クラスタ 4	1	1
クラスタ 5	4	2
クラスタ 6	0	2
クラスタ 7	7	0

以上，2 つの評価基準によって，モチーフによるクラスタリングの有効性を示すことができた．

5. おわりに

本稿では，モチーフに改良を加えた上で波形データをクラスタリングする手法を提案し，K-means によるクラスタリングと比較して，その有効性および課題について述べた．時系列波形データの前処理については，他にもいくつか提案されており，例えば，平野・津本らは，多重スケールマッチングにより方法を提案している [平野 03]．波形データのどの特徴に注目してどのような前処理を施せば，知識発見支援の視点から有効になるのか，検討すべき課題は多い．

今後は，モチーフに付帯するパラメータ群の設定法を考察した上で，専門家が実際に結果を評価することを取り入れ，より多くのデータで評価を進める予定である．

謝辞

本稿で題材とした慢性ウイルス性肝炎データセットを提供していただいた千葉大学病院医療情報部高林克日己医師，横井英人医師に深く感謝する．また，本研究は，文部科学省科学研究費補助金特定領域研究 (13131205)「メタ学習機構に基づくアクティブマイニング」の助成によるものである．

参考文献

[Ohsaki02] M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi: ”A Rule Discovery Support System for Sequential Medical Data, – In the Case Study of a Chronic Hepatitis Dataset –”, Int’l Workshop on Active Mining (AM-2002) in the IEEE Int’l Conf. on Data Mining (ICDM’02), Maebashi, Japan, pp.97–102(2002).

[Lin02] Lin, J., Keogh, E., Patel, P., and Lonardi, S.: ”Finding Motifs in Time Series”,KDD2002, pp.23-26(2002).

[平野 03] 平野章二, 津本周作:” 多重スケールマッチングにより導出される類似度の性質”, 人工知能学会 第 60 回知識ベースシステム研究会 (2003)