階層的クラスタリングを用いた時系列データの分割†

高橋 勝稔*¹・馬野 元秀*²・藤本 典幸*³

長期間の時系列データは傾向や特徴の推移により理解されていると考えられる。そのために、傾向や特徴が変化している時刻で時系列を分割する必要がある。例えば、統計学の異常検知や変化検知では、異常や変化の度合いを確率分布を用いて表す方法などが提案されている。我々は、階層的クラスタリングを用いて時系列データを分割する手法を提案する。本手法では、分割された各期間をクラスタとみなし、クラスタ間の類似性が大きいものをまとめていく。類似性にはクラスタの値の大きさ、値の変化、振動を定義し、これらの重み付き和で求めた距離を基に決定する。しかし、重みを固定するとうまくまとまらないので各類似性の大小関係やクラスタの大きさなどを用いて重みを動的に変化させる。また、振動している部分を1つのクラスタにまとめるために、小さいクラスタをまとまりやすくするパラメータや、はずれ値などを含む小さいクラスタの前後のクラスタが似ていればまとめる手法も提案する。この手法を実際の時系列データに適用し、その有用性を調べる。

キーワード：時系列データ、階層的クラスタリング、時系列データの分割

1. はじめに

時系列データとは、時間順に並んだ数値データのことで、身近な例としては毎日の気温の変化や株価の変化、学生の試験の点数の推移などがある。人間は、時系列データの数値や、それを視覚的に表したグラフなどの情報に対して優れた処理能力を持っている。例えば、長期間の時系列データを理解するとき、「全体的には少し増加しているが、はじめの方は少し下がっていて終わりの方は振動が多い」などのように、全体的な傾向といくつかの期間の傾向や特徴の推移により理解していると考えられる。これらの期間、あらかじめ決められたものもあるが（月の初旬、中旬、下旬など）、データの傾向や特徴によって変わることも多い。このような期間の分割を計算機で行うことは人間と機械のコミュニケーションという観点で重要である。

長期間の時系列データに対して傾向や特徴が変化する時刻を決めるために、例えば、統計学における異常検知や変化検知の分野では、異常や変化の度合いを確率分布を用いて表す方法などが提案されている [1]。また、周期性のある傾向や特徴の類似性を調べるときにはスペクトル解析を用いることが多い [2]。他にも、隠れマルコフモデル（HMM: Hidden Markov Model）を利用した期間の分割方法なども提案されている [3, 4]。しかし、これらの方法は、時系列データに周期性やマルコフ性があることを仮定しているために、適用できない時系列データも少

なくない。また、値の変化を用いた分割法 [5] も提案されているが、これは、ゆるやかな線形性を考え、線形でなくなる時刻で分割する手法であるので、振動が大きい部分はまとまらず、細かく分割され不自然な結果となる。

本論文では、期間毎の値の大きさと値の変化と振動を用いて距離を定義し、距離を基にして隣り合う期間で類似している部分をまとめることにより時系列データの分割を行う手法を提案する。

まず、時系列データの時間軸上で隣り合うデータを結んだ線分の1つずつを1つのクラスタ（期間を表している）とする。そして、各クラスタの値の大きさ、値の変化、振動を基にして隣り合うクラスタの傾向が類似しているものから順にまとめていく。クラスタの類似性は3つの類似性の重み付き和を用いるが、重みを固定すると人間の感覚に合わない結果になる。そこで、各類似性の大小関係やクラスタの大きさなどを用いて重みを調整する手法を提案する。また、はずれ値などを含む小さいクラスタの前後のクラスタが似ていればまとめる手法も提案する。

クラスタリングには、大別すると階層的クラスタリングと非階層的クラスタリングの2つがある [6]。階層的クラスタリングはクラスタがまとまっていく様子を見ることができ、最終的なクラスタ数を調整しやすいので、提案手法では階層的クラスタリングを用いる。階層的クラスタリングでは非階層的クラスタリングと比べて計算量が多くなるが、隣り合うクラスタ（または小さいクラスタを挟んだクラスタ）しかまとめないで、すべてのクラスタ間の距離を調べる必要はなく、通常の階層的クラスタリングと比べて計算量は少なくすむ。そして、提案手法を実際の時系列データに適用し、その有用性を調べる。

2. 時系列データの階層的クラスタリング

本論文では、（時間軸上で）隣り合う時系列データの傾向が類似しているものを1つにまとめていくことにより少数のクラスタを求める手法を提案する。クラスタの初期値は、例えば、

† Partition of Time Series in Temporal Axis Using Hierarchical Clustering
Katsutoshi TAKAHASHI, Motohide UMANO,
and Noriyuki FUJIMOTO

*1 大阪府立大学大学院理学系研究科
Graduate School of Science, Osaka Prefecture University
*2 日立造船株式会社 技術研究所, 大阪府立大学 名誉教授
Technical Research Institute, Hitachi Zosen Corporation and Professor
Emeritus of Osaka Prefecture University
*3 大阪府立大学大学院工学系研究科
Graduate School of Engineering, Osaka Prefecture University

図1 クラスタリングの初期状態と終了状態

図1の細い実線で囲った期間のように隣り合うデータを結ぶ線分となる。最終的に隣り合うクラスタがまとまっていくと図1では太い実線で囲った大きい3つのクラスタとなり、それらのクラスタの境界の点で分割することにより3つの期間に分けることができる。

まず、時系列データ全体をデータ点 (t_i, x_i) の集合 $\{(t_i, x_i) \mid i = 1, 2, \dots, n\}$ と表す。そして、 (t_i, x_i) と (t_{i+1}, x_{i+1}) を結ぶ線分を $\langle x_i, x_{i+1} \rangle$ と表すと、隣り合うデータ点を結ぶすべての線分 $\langle x_1, x_2 \rangle, \langle x_2, x_3 \rangle, \dots, \langle x_{n-1}, x_n \rangle$ がクラスタリングの始めの状態となる。そして、傾向が似ているクラスタから順にまとめていく。このとき、通常のクラスタリングとは異なり、隣り合うクラスタのみをまとめていく。クラスタ内に複数の線分がある場合、線分 $\langle x_i, x_{i+1} \rangle, \langle x_{i+1}, x_{i+2} \rangle, \dots, \langle x_{i+k-1}, x_{i+k} \rangle$ からなるクラスタを $\langle x_i, x_{i+1}, \dots, x_{i+k} \rangle$ または $\langle x_i, x_{i+k} \rangle$ で表すことにする。

(1) クラスタの類似性

クラスタが似ているということをどのように定義するかについてはさまざまなものが考えられる。ここでは値の大きさが似ている、値の変化の仕方が似ている、値の振動の仕方が似ているの3つの類似性を考える。

(a) 値の大きさによる類似性

隣り合うクラスタ c と c' の値の大きさが似ているものは類似性が高いと考えられる。クラスタの値の大きさを表す量として、クラスタ内のデータの平均値を正規化したものを用いる。

k 個の線分 $\langle x_i, x_{i+1}, \dots, x_{i+k} \rangle$ からなるクラスタ c の値の大きさを表す量 $v(c)$ は平均値 $m(c) = (x_i + x_{i+1} + \dots + x_{i+k}) / (k+1)$ を求め、それをグラフに描くときの縦軸の最大値 $x_{\max}$ と最小値 $x_{\min}$ を用いて正規化し、

$$v(c) = \frac{m(c) - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

とする。他の類似性も考慮する必要があるので、0 から 1 の範囲にするために、時系列のグラフを描くときの $x_{\max}$ と $x_{\min}$ を用いている (x_i の最大値と最小値でもよい)。

クラスタをまとめる毎に $v(c)$ は変わるが、 $x_i + x_{i+1} + \dots + x_{i+k}$ と k を憶えておくことにより逐次的に更新できる。

隣り合うクラスタ c と c' の値の大きさを表す量の差は

$d_v(c, c') = |v(c) - v(c')|$ となる。この値はデータの値が $x_{\max}$ と $x_{\min}$ の範囲に収まっていれば 0 と 1 の間の値となる。

(b) 値の変化の仕方による類似性

隣り合うクラスタ c と c' の値の変化の仕方が似ているものは類似性が高いと考えられる。クラスタの値の変化を表す量として、クラスタ内のデータを用いて最小二乗法で求めた直線の時間軸からの角度を用いる。傾きも考えられるが、傾きの範囲は $-\infty$ から $+\infty$ となる (実際は負の大きい値から正の大きい値となる) ので、角度の方が取り扱いやすいと考えられる。角度は、グラフに描いたときの t_i と x_i の単位あたりの長さの比により変わるので、 k 個の線分 $\langle x_i, x_{i+1}, \dots, x_{i+k} \rangle$ からなるクラスタ c の値の変化を表す量は、グラフの縦軸の値 $x_{\max}$ と $x_{\min}$ 、横軸の値 t_n と t_1 、グラフの縦横比 r を用いて、

$$\theta(c) = \arctan \left(\frac{\text{Cov}(T, X)}{\sigma_t^2} \cdot \frac{t_n - t_1}{x_{\max} - x_{\min}} \cdot r \right) / \pi \quad (2)$$

となる。ここで、 $\text{Cov}(T, X)$ は $T = \{t_i, t_{i+1}, \dots, t_{i+k}\}$ と $X = \{x_i, x_{i+1}, \dots, x_{i+k}\}$ の共分散で、 σ_t^2 は T の分散を表し、 $\text{Cov}(T, X) / \sigma_t^2$ は傾きとなる。縦横比が決まっていなかった場合には、 $r = 0.5$ を用いる。角度の範囲は $-\frac{\pi}{2}$ から $\frac{\pi}{2}$ なので $\theta(c)$ の値域は $-\frac{1}{2}$ から $\frac{1}{2}$ となる。 $\theta(c)$ は T の和と自乗和、 X の和と自乗和、 k を憶えておくことにより逐次的に更新できる。

隣り合うクラスタ c と c' の値の変化を表す量の差は $d_\theta(c, c') = |\theta(c) - \theta(c')|$ となる。この値は 0 と 1 の間の値となる。

(c) 振動の仕方による類似性

隣り合うクラスタ c と c' の振動の仕方が似ているものは類似性が高いと考えられる。クラスタの振動の仕方を表す量として、ここでは値の散らばり具合を考え、値の標準偏差を用いることにした。値の増減を用いた振動回数を考慮した方法も研究されている [7] が、本論文では用いなかった。

k 個の線分 $\langle x_i, x_{i+1}, \dots, x_{i+k} \rangle$ からなるクラスタ c の標準偏差は $\sigma'(c) = \sqrt{\frac{1}{k+1} \sum_{j=i}^{i+k} (x_j - m)^2}$ となる。ただし、 $m = \frac{1}{k+1} \sum_{j=i}^{i+k} x_j$ である。値を 0 から 1 にするために、すべてのクラスタにおける標準偏差 $\sigma'(*) = \{\sigma'(c_1), \dots, \sigma'(c_k)\}$ の最大値から最小値を引いた値で割り、

$$\sigma(c) = \frac{\sigma'(c) - \min\{\sigma'(*)\}}{\max\{\sigma'(*)\} - \min\{\sigma'(*)\}} \quad (3)$$

とする。 $\sigma(c)$ は $\{x_i, x_{i+1}, \dots, x_{i+k}\}$ の和と自乗和、 k を憶えておくことにより逐次的に更新できる。

隣り合うクラスタ c と c' の振動を表す量の差は $d_\sigma(c, c') = |\sigma(c) - \sigma(c')|$ となる。この値は 0 と 1 の間の値となる。

(d) クラスタの全体的な類似性

隣り合うクラスタ c と c' の値の大きさを表す量の差 $d_v(c, c')$ と変化を表す量の差 $d_\theta(c, c')$ と振動を表す量の差 $d_\sigma(c, c')$ を組み合わせることにより、隣り合う 2 つのクラスタ間の全体的な差 $d_{all}(c, c')$ を求める。重み w_1 と w_2 と w_3 ($w_1 + w_2 + w_3 = 1$) を用いて、

図2 値の大きさを重視したクラスタリング (その1)

図4 値の変化を重視したクラスタリング (その1)

図3 値の大きさを重視したクラスタリング (その2)

図5 値の変化を重視したクラスタリング (その2)

$$d_{all}(c, c') = w_1 d_v(c, c') + w_2 d_\theta(c, c') + w_3 d_\sigma(c, c') \quad (4)$$

とする。この値は2つのクラスタ間の距離と考えられるので、小さいほど2つのクラスタは似ていることになる。

(2) 全体的な差の重みの決定

次に、隣り合う2つのクラスタ間の全体的な差を求めるときに用いる重み w_i について考える。まず重みを固定したときの問題点を挙げ、次に、クラスタの時間軸上の大きさ（幅）を用いて重みをクラスタ間毎に変更する手法について述べる。

(a) 重みを固定したときの問題点

まず、クラスタの値の大きさが重要であるとして、 w_1 の重みを大きくしたときを考える。例えば、 $w_1 = 0.7, w_2 = 0.25, w_3 = 0.05$ とし、図2の時系列データを3つのクラスタに分割する（クラスタの終了条件は(6)で述べる）と、縦の実線を境界とするクラスタが得られる。この結果はうまくクラスタリングされていると考えられる。

次に、同じ重みで図3の時系列データを2つのクラスタに分割すると、縦の実線を境界とするクラスタが得られる。これは、値の大きさを重視したので値の小さい前半と値の大きい後半の2つに分割されている。しかし、この場合は $t = 4$ で分割されるのが自然であると思われる。これは、 $t = 0$ から4までは減少しており、 $t = 4$ から15までは増加しているという特徴を優先して考えていることになる。

そこで、クラスタの変化の重み w_2 を大きくして、 $w_1 = 0.25, w_2 = 0.7, w_3 = 0.05$ にすると、図3は図4のようにうまくいくが、図2が図5のように分割されてしまう。これは $t = 0$ から5と $t = 10$ から15までは初期クラスタの変化の仕方が似

ているので序盤で1つのクラスタとなるが、 $t = 5$ から10の振動している部分は隣り合う初期クラスタの値の変化が大きく異なるので1つのクラスタにならず、クラスタリングの終盤で $t = 0$ から5のクラスタと $t = 5$ から6がまとまり、そのクラスタと $t = 6$ から7がまとまる、というように順にまとまるからである。これでは $t = 5$ から10の振動している部分が1つのクラスタにならない。

(b) クラスタ毎の重みの変更

これらの問題を解決するために、クラスタの大きさにより重み w_1, w_2, w_3 の値を変えることを考える。

● クラスタが小さいとき

図2の $t = 5$ から10の振動している部分は1つのクラスタとしてまとめたい。この部分が1つにまとまるためには、クラスタが小さいときに値の変化よりも大きさを重視した重みにすればよい。また、振動している部分は変化による類似性が低いのであまり考慮しない方がよい。しかし、それでは図2の $t = 0$ から5のように変化による類似性が高く値による類似性が低い場合に1つにまとまらない。

そこで、クラスタの値の大きさと変化の小さい方を重視した重み付けにする（同じ場合は値を重視する）。すなわち、図2の $t = 0$ から1と $t = 1$ から2のように変化の仕方が似ていると値の大きさは異なり、 $t = 5$ から6と $t = 6$ から7のように値の大きさが似ていると変化の仕方は異なるので、似ている方の重みを大きく、似ていない方の重みを小さくする（極端な場合は0とする）。本論文では3つの重みを0.94, 0.04, 0.02とし、値の大きさの差 $d_v(c, c')$ と変化の差 $d_\theta(c, c')$ の小さい方の重

図6 関数 $S(l; a_i, b_i)$

みを 0.94, 大きい方の重みを 0.04 とし, 振動の重みを 0.02 とする. すなわち, $d_v(c, c') \leq d_\theta(c, c')$ のときには, クラスタが小さい時の値の重みを $W_1^S = 0.94$, 変化の重みを $W_2^S = 0.04$, 振動の重みを $W_3^S = 0.02$ とし, $d_v(c, c') > d_\theta(c, c')$ のときには, $W_1^S = 0.04$, $W_2^S = 0.94$, $W_3^S = 0.02$ とする. なお, これは, 値の大きさの差 $d_v(c, c')$ と変化の差 $d_\theta(c, c')$ だけを考えると, 順序付き加重平均 (OWA: ordered weighted average) [8] となっている.

● クラスタが大きいとき

クラスタが大きいときは, クラスタの値の大きさより変化の方を重視したいので変化の重み w_2 を大きくする. また, 変化の仕方が似ていると, 一般には値は異なるので, 値の重みを小さくする (極端な場合は 0 にする). 例えば, クラスタが大きいときの値の重みを $W_1^L = 0$, 変化の重みを $W_2^L = 0.9$, 振動の重みを $W_3^L = 0.1$ とする.

一般の場合は, クラスタが大きいときの重み W_i^L と小さいときの重み W_i^S の重み付き平均を用いる. そのために, クラスタの大きさを定義する. クラスタ $c = \langle x_i, x_{i+1}, \dots, x_{i+k} \rangle$ の大きさ $l(c)$ は, t_1 から t_n を 1 としたときのクラスタの期間の割合で定義し,

$$l(c) = (t_{i+k} - t_i) / (t_n - t_1) \quad (5)$$

とする. 隣り合うクラスタ c と c' の大きさ $l(c)$ と $l(c')$ の小さい方を l として, $d_{all}(c, c')$ を求める式 (4) の重み w_i を

$$w_i = S(l; a_i, b_i) \cdot W_i^L + (1 - S(l; a_i, b_i)) \cdot W_i^S \quad (6)$$

と定義する. ここで, $S(l; a_i, b_i)$ は Zadeh の定義した標準関数の S 関数 [9] で, 図 6 のようになる. $l \geq b_i$ すなわちクラスタが十分に大きいときは $S(l; a_i, b_i) = 1$ で $w_i = W_i^L$ となり, $l < a_i$ すなわちクラスタが十分に小さいときは $w_i = W_i^S$ となる. $a_i \leq l < b_i$ のときは W_i^L と W_i^S の $S(l; a_i, b_i)$ による重み付き平均となる. a_i と b_i は値の大きさ, 変化, 振動に対して異なる値を設定できるが, 本論文ではすべて同じ値 ($a = 0.1$, $b = 0.2$) を用いた. なお, $w_1 + w_2 + w_3 \neq 1$ となる場合は $w_i = w_i / (w_1 + w_2 + w_3)$ とし, 正規化を行っている.

この手法で分割すると図 2 と図 4 のようにうまく分割できるようになる. 値の変化を用いた時系列データの分割法として文献 [5] などの方法があるが, 期間をゆるやかな線形で捉えてい

るため, 振動の多い部分は細かく分割されやすく, 図 2 のように振動している部分をまとめた分割は得られない. また, 周波数解析などで特定の周波数などを調べるよりも簡単で高速かつ人間にとってわかりやすい特徴量を用いた分割となっている.

(3) クラスタの大きさによる補正

これまでの手法を実際の時系列データに適用する. 図 7 はごみ焼却発電施設の, あるセンサから得られた 1 分毎の時系列データを表しており, データ数は 481 である. 本手法を用いて分割すると, クラスタが 9 個のときは縦の実線が境界となる. この場合, $t = 300$ 付近の急激に変化している部分は 2 つの小さいクラスタとして残っている. しかし, このような場合は, 全体として振動している部分 (あるいは山型) として解釈する方が自然である.

そうするためには, 小さいクラスタほどまとまりやすくすればよい. そこで, 隣り合うクラスタ c と c' の大きさ $l(c)$ と $l(c')$ の値の小さい方を l として, l が小さいときはクラスタの全体的な差 d_{all} が小さくなり, l が大きいときは d_{all} が大きくなるように, d_{All} を

$$d_{All}(c, c') = d_{all}(c, c') \cdot (1 + k \cdot l) \quad (7)$$

とする. ここで, k は正のパラメータである. $l = 0$ のときは, $d_{All}(c, c') = d_{all}(c, c')$ となり, $l = 1$ のときは, $d_{All}(c, c') = d_{all}(c, c') \cdot (1 + k)$ となる. これにより, 大きいクラスタは小さいクラスタより d_{All} の値が大きくなるので, 先に小さいクラスタがまとまりやすくなる. k の値を大きくすると小さいクラスタがまとまりやすくなるが, 値を大きくし過ぎると, 同じ大きさのクラスタとなってしまう.

図7 ごみ焼却発電施設の時系列データ (その 1)

図8 クラスタの大きさによる補正を行った結果

図9 1つ越えたクラスタが似ている例

図10 1つ越えたクラスタををまとめた例

$k = 50$ として図7をクラスタリングした結果は図8（クラス数9で図7と同じ）のようになり、 $t = 120$ 付近のクラスタよりも先に $t = 300$ 付近が1つにまとまっている。

(4) 1つ越えたクラスタ間の類似性

図9において、 $t = 0$ から10のクラスタを c_1 とし、 $t = 10$ から12のクラスタを c_2 とし、 $t = 12$ から19のクラスタを c_3 とすると、 c_1 と c_3 は似ているが、 c_2 は c_1 と c_3 とともに似ていなくて、 $l(c_2)$ は小さい。このような場合、クラスタ c_2 を無視して、 c_1 から c_3 までを1つのクラスタにまとめたい。しかし、隣り合うクラスタしかまとまらないので c_1 と c_2 が先にまとまってしまうと、 c_3 とは似ていないクラスタになる。

そこで、1つ越えたクラスタ c_i と c_{i+2} の全体的な差 $d'_{All}(c_i, c_{i+2})$ も計算し、隣り合うクラスタ c_i と c_{i+1} の全体的な差 $d_{All}(c_i, c_{i+1})$ と合わせて比較する。このとき、 $l(c_{i+1})$ が $l(c_i)$ や $l(c_{i+2})$ と比べて小さいときだけ1つ越えたクラスタをまとめるべきなので、 c_i と c_{i+2} の類似度を

$$d'_{All}(c_i, c_{i+2}) = d_{All}(c_i, c_{i+2}) \cdot \left(1 + \frac{l(c_{i+1})}{\min(l(c_i), l(c_{i+2}))}\right) \quad (8)$$

と定義する。 d'_{All} は $l(c_{i+1})$ が $l(c_i)$ と $l(c_{i+2})$ より小さいときに値が小さくなり、大きいときは値は大きくなる。ただし、 $d_{All}(c_i, c_{i+2})$ が小さい場合には、 $\left(1 + \frac{l(c_{i+1})}{\min(l(c_i), l(c_{i+2}))}\right)$ の値に関係なく $d'_{All}(c_i, c_{i+2})$ の値が小さくなってしまふので、 $\frac{l(c_{i+1})}{\min(l(c_i), l(c_{i+2}))} < \theta_w$ の場合にのみクラスタリングの対象とする。ここでは $\theta_w = 0.5$ とした。さらに、 c_i と c_{i+2} が似ていても c_{i+1} の始めと終わりのデータ（それぞれ x_a と x_b とする）が縦軸方向に離れていると c_i と c_{i+2} をまとめることはできないと考え、 $\frac{|x_a - x_b|}{x_{\max} - x_{\min}} < \theta_h$ の場合にのみクラスタリングの対象と

図11 クラスタリング途中の様子（クラス数77）

図12 クラスタリングの途中の様子（クラス数11）

する。ここでは $\theta_h = 0.2$ とした。

なお、1つ越えたクラスタがまとまると、間の小さいクラスタは以降のクラスタリングでは用いない。

1つ越えたクラスタ間の類似性を考慮したことにより図9は、次に c_1 と c_3 がまとまり、 $t = 0$ から19のクラスタと $t = 19$ から30の2つのクラスタになる。1つ越えたクラスタがまとまった場合、無視された間のクラスタ（ $t = 10$ から12）は図10のように縦の高さが低い破線で示される。

(5) 実際の時系列データへの適用（その1）

この手法を(3)の時系列データに適用する。なお、各パラメータは、これまでと同様に、クラスタの大きさによる補正のパラメータ $k = 50$ 、1つ越えたクラスタ間の類似性のしきい値 $\theta_w = 0.5$ 、 $\theta_h = 0.2$ を用いた。 $d_{All}(c_i, c_{i+1})$ と $d'_{All}(c_i, c_{i+2})$ が最も小さいもの（最も小さいものが複数個あるときは最も左のもの）から順に1つずつまとめていくと、クラスタ数が77のときは図11のようになる。この時、 $t = 213$ から214の小さいクラスタの前後のクラスタ（それぞれ $t = 206$ から213と $t = 214$ から217）の値の大きさが似ているので、初めて1つ越えたクラスタがまとまる。なお、図11では重なって1本の破線に見えるが、図10と同様に2本の破線（ $t = 213$ と $t = 214$ ）がある。

さらにまとめていくと、クラスタ数が11の時は図12のようになる。このとき、無視されたクラスタは $t = 213$ から214と、 $t = 464$ から465と、 $t = 459$ から460の3つとなる。ここで、図12だけを見ると、 $t = 213$ から214のクラスタは前後のクラスタ（ $t = 206$ から213と $t = 214$ から288）が似ていて、1つ越えたクラスタリングが行われたと考えてしまうが、

すでに図 11 で 1 つ越えたクラスタリングが行われている。このように 1 つ越えたクラスタリングが行われた時期や、そのときの前後のクラスタの状態はクラスタリング結果の図では分からないことに注意する必要がある。

(6) 終了条件

階層的クラスタリングはクラスタが 1 つになるまでまとめることができるので、適切と思われるクラスタ数で終了する必要がある。クラスタリングの終了条件は、まず全期間における大きいクラスタの占める割合を

$$\delta_l = \frac{\sum_c S(l(c); a, b) \cdot l(c)}{t_n - 1} \quad (9)$$

で定義し（進捗率と呼ぶことにする）。この値がしきい値 θ_l を超えたときとする。 $S(l(c); a, b)$ は式 (6) と同じ関数で、パラメータはここでは $a = 0.1, b = 0.2$ としている。これは、すべてのクラスタの大きさが全期間の 10% 以下であれば $S(l(c); a, b)$ が 0 で進捗率が 0 となり、すべてのクラスタの大きさが全期間の 20% 以上であれば $S(l(c); a, b)$ がすべて 1 で進捗率は 1 となる。これに加えて、クラスタの個数がしきい値 θ_n 以下となるときとしている。今回は進捗率のしきい値 θ_l を 0.8、個数のしきい値 θ_n を 8 とした。なお、個数を優先させたい場合は、 θ_l を 0 とし、進捗率を優先させたい場合は、 θ_n をデータ数とすればよい。

図 12 からさらに進めると、クラスタ数が 8 の時は図 13 のようになる。この時の進捗率は 0.601 である。さらにクラスタリングが進むと、クラスタ数が 7 で進捗率は 0.706 となり、クラスタ数が 6 で進捗率は 0.842 となる。このとき、進捗率と個数のしきい値の両方の条件を満たすので、クラスタリングが終了

図 13 クラスタリングの途中の様子（クラスタ数 8）

図 14 クラスタリング結果（クラスタ数 6）

する。終了状態は図 14 となり、各クラスタの位置、振動、全体的な傾向は 1 つ目 ($t = 0$ から 18) が高い位置で少し振動しながら中ぐらい減少、2 つ目 ($t = 18$ から 125) がやや高い位置で少し振動しながらほぼ一定、3 つ目 ($t = 125$ から 206) が中ぐらいの位置で振動しながらほぼ一定、4 つ目 ($t = 206$ から 288) が中ぐらいの位置で少し振動しながら中ぐらい増加、5 つ目 ($t = 288$ から 316) が高い位置であまり振動はなく大きく上に凸、6 つ目 ($t = 316$ から 480) がやや高い位置で大きく振動しながらほぼ一定となる。

今回は $a = 0.1, b = 0.2$ としたので、例えば同じ大きさのクラスタであれば 5 個のときに進捗率が 1 となる。8 個くらいで終了したいときは $a = 0, b = 0.125$ などとするのがよい（ちょうど 8 個で終了したい場合は a, b をより小さい値にして個数で指定すればよい）。また、終了条件を進捗率とクラスタの個数のどちらかの成立とすることも考えられる。

(7) 実際の時系列データへの適用（その 2）

他の実際の時系列データに対してクラスタリングを行った結果を図 15 に示す。クラスタリングの終了条件は (5) と同じにした。

図 15 では、順に（全体的な傾向はすべてほぼ一定で）高い位置で少し振動している部分と、低い位置で中ぐらい振動している部分と、時々振動しているが高い位置ではほぼ一定である部分と、大きく振動している部分の 4 つに分けられる。 $t = 318$ から 319 や $t = 451$ から 454 などのクラスタは前後の傾向が似ているので 1 つ越えたクラスタリングが行われた。

$t = 260$ から 274 も同様に 1 つ越えたクラスタリングが行われたが、無視されたクラスタが他の場合とくらべて大きいものが、 $t = 223$ から 260 と $t = 274$ から 370 の大きな前後のクラスタが似ているので 1 つ越えてまとめたものである。最終的には、 $t = 127$ から 370 のクラスタにおいて、この部分は大きく減少している外れ値と考えられる。しかし、その大きさは 17 で全期間の大きさ 480 と比べて大きく、値も 39.7 から 9.3 まで減少しているために（その後 36 まで増加するが）、クラスタの境界と考えることもできる。どちらにするかはパラメータに依存し、クラスタの境界にするには 1 つ越えたクラスタ間の類似性を計算するときのしきい値 θ_w の値を下げればよい。しかし、その他の 1 つ越えたクラスタもまともにならなくなってしまう。

図 15 ごみ焼却発電施設の時系列データのクラスタリング（その 2）

図 16 ごみ焼却発電施設の時系列データのクラスタリング (その3)

図 17 終了状態からさらにクラスタリングを行った結果

う。より一般的にはクラスタリングの時期によって θ_w を動的に変更する必要があると思われる (今回は考えない)。

別の例を図 16 に示す (パラメータは同じ)。図 16 では、6 つのクラスタに分けられた。 $t = 133$ から 144 の小さいクラスタは単調減少で変化も急激なために、隣のクラスタとまとまることなくクラスタリングが終了している。終了状態の図 16 からさらにクラスタリングを行う (クラスタ数 4) と図 17 のように小さいクラスタがなくなり、高い位置で振動している部分と、減少している部分と、中ぐらいの位置で振動している部分と、低い位置であまり振動していない部分に分けられる。これを考えると終了条件は今回の手法の他にも、小さいクラスタがなくなったときなども考えられる。その場合には、2 つめのクラスタは単調減少ではなくなり、4 つめのクラスタには振動している部分が少し含まれてしまう。

小さいクラスタをなくしたい場合、式 (7) の小さいクラスタがまとまりやすくなるパラメータ k を大きくすることも考えられる。しかし、大きくしすぎるとすべてのクラスタリングでクラスタの大きさを重視し、同じ大きさのクラスタとなってしまう。改良手法としては、 k も θ_w と同様に動的に設定することが考えられる。

3. おわりに

本論文では、人間が長期間の時系列データを理解するとき、複数の期間に分割してそれらの傾向を理解していると考え、時系列データを傾向に基づいて分割する手法について述べた。期間をクラスタとみなし、各クラスタの値の大きさと変化と振動に基づいて距離を求め、階層的クラスタリングを行っ

た。距離の計算では、クラスタの大きさと傾向によって重みを動的に変化させている。また、外れ値のような前後と比べて期間が短く傾向が異なるクラスタについて、その前後のクラスタが、似ていればまとまるように 1 つ越えたクラスタの類似性も考慮した。

本手法を用いると、長期間の時系列データからよく似た傾向の続いている部分を取り出すことができ、これに基づいて、傾向の持続時間や変化する時期を調べることで、実際のセンサーデータの特徴を知ることができる。今後は応用を通してクラスタリングの終了条件や動的なパラメータの設定を検討していきたい。

参考文献

- [1] 井手剛, 杉山将: 異常検知と変化検知, 機械学習プロフェッショナルシリーズ, 講談社, 2015.
- [2] 日野幹雄: スペクトル解析, 朝倉書店, 2010.
- [3] 白山晋: “時系列速度データからのパターン抽出法,” 第 29 回数値流体力学シンポジウム, C09-1, 2015.
- [4] 松原靖子, 櫻井保志, C. Faloutsos: “大規模時系列データからの特徴自動抽出,” 第 6 回データ工学と情報マネジメントに関するフォーラム, D4-2, 2014.
- [5] J. Kacprzyk, A. Wilbik, and S. Zadrozny: “On Some Types of Linguistic Summaries of Time Series,” *Proc. of the 3rd Int. IEEE Conf. on Intelligent Systems*, pp. 373-378, 2006.
- [6] 宮本定明: クラスタ分析入門 —ファジィクラスタリングの理論と応用, 森北出版, 1999.
- [7] 馬野元秀, 岡村光洋, 瀬田和久: “全体的傾向と局所的特徴に基づく時系列データの言葉による表現 —増減回数の拡張による振動の表現—,” 第 24 回ファジィシステムシンポジウム, pp. 774-779, 2008.
- [8] R. R. Yager: “OWA Aggregation over a Continuous Interval Argument with Applications to Decision Making,” *IEEE Trans. Syst., Man, Cybern.*, Vol.34, pp. 1952-1963, 2004.
- [9] L. A. Zadeh: “Fuzzy Logic and Approximate Reasoning,” *Synthese*, Vol.30, pp. 407-428, 1975.

(2018 年 7 月 2 日 受付)

(2018 年 11 月 20 日 採録)

[問い合わせ先]

〒599-8531 大阪府堺市中区学園町 1-1

大阪府立大学 大学院理学系研究科 情報数理科学専攻

高橋 勝稔

TEL: 072-981-8792

FAX: 072-981-8792

E-mail: katsu_t1984@yahoo.co.jp

著者紹介

たかはし かつとし
高橋 勝稔 [学生会員]

2009 年大阪府立大学 総合科学部 数理・情報科学科卒業。現在、同大学 大学院理学系研究科 情報数理科学専攻 博士後期課程在学中。時系列データの言語化、クラスタリング、言葉による検索などの研究に従事。

うまの もとひで
馬野 元秀 [名誉会員]

1974 年 3 月, 大阪大学 基礎工学部 情報工学科卒業. 1979 年 3 月, 同大学院基礎工学研究科 物理系専攻 情報工学分野後期課程修了, 工学博士. 同年 4 月, 岡山理科大学 理学部応用数学科 講師. 1985 年 4 月, 大阪大学 大型計算機センター 助手. 同センター 講師, 助教授を経て, 1991 年 1 月, 同大学 工学部精密工学教室 助教授. 1993 年 10 月, 同大学 基礎工学部 システム工学科 助教授. 1996 年 4 月, 大阪府立大学 総合科学部 数理・情報科学講座 教授. 2005 年, 改組により大学院理学系研究科 情報数理科学専攻 教授. 2016 年 4 月, 同大学 工学研究科 知能情報工学分野 教授. 2017 年 3 月, 大阪府立大学 定年退職. 名誉教授. 現在, 日立造船株式会社 事業企画・技術開発本部 技術研究所 ものづくり基盤研究センター 技術顧問. ファジィ集合論の応用, 特に, プログラミング言語, データベース, 知識情報処理への応用に関する研究に従事. 情報処理学会, 人工知能学会, システム制御情報学会, 計測自動制御学会, ACM, IEEE の各会員.

ふじもと のりゆき
藤本 典幸 [正会員]

1997 年 3 月大阪大学大学院基礎工学研究科物理系専攻情報工学分野博士後期課程単位取得退学. 同年 4 月大阪大学大学院基礎工学研究科助手. 2002 年 4 月大阪大学大学院情報科学研究科助教授, 2007 年 4 月同研究科准教授, 2008 年 4 月大阪府立大学大学院理学系研究科教授. 2016 年 4 月大阪府立大学大学院工学研究科教授となり現在に至る. 情報科学, 特に並列処理のソフトウェアに関する研究に従事. 工学博士. 2003 年度第 3 回船井情報科学奨励賞ほかを受賞, IEEE, ACM, 情報処理学会, 電子情報通信学会などの会員.

Partition of Time Series in Temporal Axis Using Hierarchical Clustering

by

Katsutoshi TAKAHASHI, Motohide UMANO, and Noriyuki FUJIMOTO

Abstract:

Long-term time series data can be understood according to transitions of trends and features. We must partition time series into several terms of intervals of different trends and features. In statistics, change detection expresses it as a degree of change using probability distributions. In this paper, we propose a method to partition time series using a hierarchical clustering. First, we have clusters of one line segment connecting adjacent data. Then, we merge two adjacent similar clusters into one with a total similarity calculated by the weighted average of three similarities of values, change of values and oscillations. However, the partition results with the fixed weights do not fit our sense. We, therefore, propose variable weights with three similarities and sizes of adjacent clusters. Furthermore, in order to exclude small clusters including outliers, we define a similarity of two clusters adjacent to the small cluster. We apply this method to actual time series and show results.

Keywords: time series, hierarchical clustering, partitioning of time series

Contact Address: **Katsutoshi TAKAHASHI**

*Graduate School of Science, Osaka Prefecture University
1-1 Gakuen, Sakai, Osaka 599-8531, Japan
TEL: +81-72-981-8792
FAX: +81-72-981-8792
E-mail: katsu_t1984@yahoo.co.jp*