2D1-2

移動平均法に基づく部分時系列クラスタリングの再解釈

Reinterpretation of Subsequence Time-series Clustering Based on Moving Average

大崎 美穂　Miho Ohsaki
片桐 滋　Shigeru Katagiri
段 純恵　Sumie Dan

同志社大学 理工学部 情報システムデザイン学科
Department of Information Systems Design, Faculty of Science and Engineering, Doshisha University

Subsequence time-series clustering (STSC) using a one-step sliding window and the k-means algorithm has a serious problem that it generates sinusoidal meaningless patterns. We discuss the similarity between STSC and moving average and examine the evidence of similarity through formulation and experiment. The results indicate that STSC works as a lowpass filter.

1. はじめに

時系列データから有益な知識を発見するニーズは高く，多くの時系列解析手法が存在する．時系列解析のアプローチは，生成メカニズムのモデル化と代表的な部分パターンの抽出に大別される．前者では，AIC 等の基準のもと，状態遷移モデルや自己回帰モデルを当てはめる．後者では，スライド窓で切り出した部分時系列をクラスタリングし，クラスタ中心を代表パターンとして抽出する方法 (STSC; Subsequence Time-Series Clustering) が一般的である．

近年，STSC の代表パターンは入力データに依存せず正弦波の形状になる，という問題が指摘された [1]．これに対し，実験的，および理論的検証が進みつつある．実験的検証 [1][2] では実データや人工データを用いた実験による現象観察を通し，理論的検証 [3][4] では現象の背景にある数学的機構の洞察を通し，問題の原理的な説明と回避策の提案がなされてきた．

本研究では，STSC と移動平均法との類似性に着目し，STSC 問題がこの標準的な信号処理法から明解な結果として説明できることを報告する．

2. STSC 問題

文献 [1] は，STSC 問題の存在を最初に指摘し，以下の処理では必ず STSC 問題が生じることを 2 種類の実験で示した (図 1 参照)．固定幅 w の窓を 1 点ごとにスライドしながら部分時系列を切り出す．部分時系列群に k 平均法を適用する．各クラスタ中心を代表的な部分パターンと見なす．

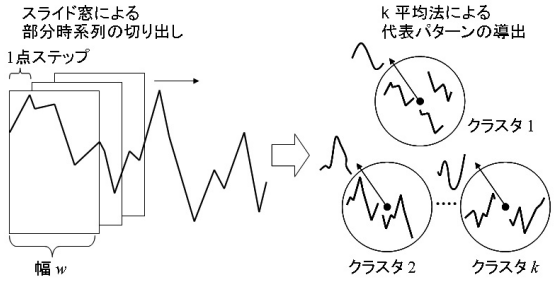

図 1: STSC の処理手続き．

連絡先: 大崎 美穂
〒 610-0321 京都府京田辺市多々羅都谷 1-3
mohsaki@mail.doshisha.ac.jp

文献 [1] の前半の実験では，w，k，初期乱数を様々に変えて株価とランダムウォークの各データに STSC を施し，クラスタ中心の再現性を調べた．また，窓をランダムに動かして部分時系列を切り出し，その後は STSC と同じ処理を行う方法 (WC; Whole Clustering) を比較対象とした．その結果，同じデータ間，異なるデータ間のどちらでも，STSC はクラスタ中心の再現性が高く WC は低かった．これは，STSC が入力データに依存せず同じクラスタ中心 (言い換えると，無意味な代表パターン) を出力することを示唆する．

後半の実験では，代表パターンの性質を調べるため，Cylinder，Bell，Funnel という 3 種類のパターンを複数連結した CBF データに $k = 3$ の STSC を適用し，得られたクラスタ中心を可視化した．その結果，クラスタ中心は 3 種類のパターンではなく正弦波に近い形状を示した．これは，STSC が正弦波様のパターンを出力する機構を持つことを示唆する．

3. 移動平均法との等価性

我々は STSC の処理機構が移動平均法に似ていると考え，両者が等価であるという仮説を立てる．なお，移動平均法とはある点を含む一定区間の平均値を求め，その点を平均値で置き換える処理である．以下では，両者を定式化し比較して仮説を検証する．計算機上で扱うことを考慮し，ここからは離散時間で考える．

k 平均法で生成される k 個のクラスタの 1 つに関し，メンバの数 n，j 番目のメンバ M_j，$j = 0, 1, ..., n-1$ とすると，クラスタ中心 C は式 (1) と表せる．STSC ではクラスタ中心とメンバは長さ w の部分時系列なので，これらを $C = (\ c[0],\ c[1],\ ...,\ c[i],\ ...,\ c[w-1]\ )$，$M_j = (\ m_j[0],\ m_j[1],\ ...,\ m_j[i],\ ...,\ m_j[w-1]\ )$ と記述する．式 (1) に基づき，C の i 番目の要素は式 (2) と表せる．

$$C = \frac{1}{n}\sum_{j=0}^{n-1} M_j \tag{1}$$

$$c[i] = \frac{1}{n}\sum_{j=0}^{n-1} m_j[i] \tag{2}$$

STSC では，1 点ずらしで原時系列から部分時系列を切り出す．このずれが部分時系列の形状差をどの程度生じるかは，原時系列の周波数成分とサンプリング周波数の関係に依存する．

しかしながら一般的に考えて，ずれが 1 点の方が数点よりも，部分時系列間の形状が類似している可能性は高い．

そこで，クラスタのメンバが 1 点づつずれた隣接部分時系列で構成されると前提する．すると，j 番目のメンバの i 番目の要素 $m_j[i]$ は，0 番目のメンバの $j+i$ 番目の要素 $m_0[j+i]$ に等しい．c を y，m_0 を x，j を $-j$ と置き直せば，式 (2) は式 (3) と表せる．この式は移動平均法の定義式，すなわち，n 点の一定区間で，移動平均法を原時系列 $x[i]$ に適用した出力 $y[i]$ と等しい．

式 (3) を z 変換することで，伝達関数が式 (4) のように求まる．さらに $z = e^{-j\omega}$ と置くことで，周波数特性が式 (5) のように求まる．式 (5) は，周波数 0 をピークとした低域通過フィルタを表す．以上より，STSC は移動平均法と等価であり，原時系列の低域通過結果 (低周波成分のみ) がクラスタ中心に現れると言える．

$$y[i] = \frac{1}{n}\sum_{j=0}^{n-1} x[i-j] \tag{3}$$

$$H[z] = \frac{1}{n}\sum_{j=0}^{n-1} z^{-j} = \frac{1}{n}\frac{1-z^{-n}}{1-z^{-1}} \tag{4}$$

$$\begin{aligned} H[e^{j\omega}] &= \frac{1}{n}\frac{1-e^{-j\omega n}}{1-e^{-j\omega}} \\ &= \frac{1}{n}\frac{e^{-j\omega n/2}}{e^{-j\omega/2}}\frac{sin(\omega n/2)}{sin(\omega/2)} \end{aligned} \tag{5}$$

4. 検証実験

前節で示した STSC と移動平均法の等価性を確認すべく，文献 [1] の前半の実験を再検証する．実験条件は文献 [1] とほぼ同じである．追加条件として，k メドイズ法を用いた STSC も調べることにした．k メドイズ法とは，クラスタ中心をメンバ平均の最近傍部分時系列とする点のみ，k 平均法と異なる．また，STSC と WC の間で対応する条件ごとに t 検定を行い，クラスタ中心の再現性の有意差を調べた．

図 2 に w，k の条件ごとの再現性を示す．左側は k 平均法を用いた STSC の結果，右側は k 平均法を用いた WC の結果である (k メドイズ法でも同様の結果を得たため，これ以降は k 平均法のみで議論する)．多くの条件で STSC が WC よりもクラスタ中心の再現性が有意に高く，文献 [1] と同様の傾向が見られた．

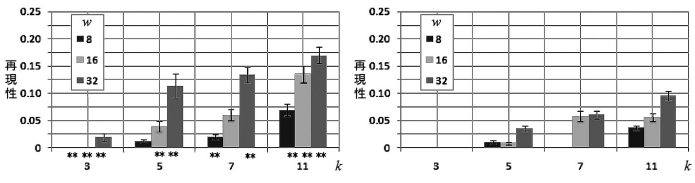

図 2: STSC(左) と WC(右) のクラスタ中心の再現性．値が大きいほど再現性が高い．

次に，クラスタが隣接部分時系列で構成されているか，および，クラスタ中心に低周波成分が現れているかを調べた．STSC と移動平均法の等価性に関し，前者は等価性の前提の成否を，後者は等価性そのものの示唆を意味する．図 3 の左側に，横軸をスライド窓の位置，縦軸をクラスタ番号としてこれらの関係を示す．本実験では窓幅 $w = 8,\ 16,\ 32$ としたが，これに比して，クラスタ 7 番，10 番は約 3000 個という多数の隣接部分時系列で構成されている．図 3 の右側に，横軸を点番号，縦軸を振幅として可視化したクラスタ中心を示す．ここでは低周波の正弦波が現れている．

以上より，実験結果は以下の様子を明確に示している．STSC におけるクラスタ中心は多数の隣接部分時系列の加算平均から成り，その結果，原時系列の低域通過成分，すなわち，正弦波様の時系列になる．

ここで STSC 問題の回避策を考察する．STSC 本来の目的は代表パターンとその位置の抽出であるため，1 点ステップのスライド窓と k 平均法の組合せが必然とは言い難い．音声信号処理で一般的な分節化のように，パワースペクトル情報でクラスタリング後，波形レベルの分節化を行う方法が考えられる．

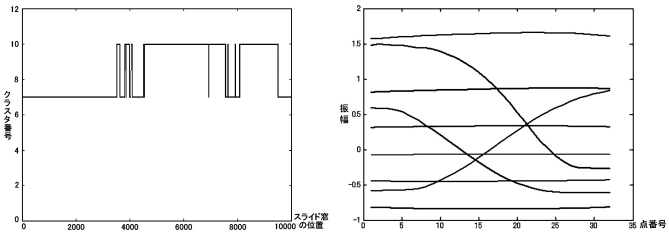

図 3: スライド窓の位置とクラスタ番号の関係 (左)．ステップ形状に近いほど近隣の部分時系列がクラスタ化されている．STSC のクラスタ中心を可視化 (右)．緩やかな変動であるほど低域通過フィルタリングの結果に近い．

5. まとめ

1 点ステップのスライド窓で切り出した部分時系列に k 平均法を適用し，クラスタ中心を代表パターンと見なす STSC には，入力データに関わらず正弦波を生じる問題があった．本研究では，STSC 問題の原理的な説明を目指し，定式化と実験によって STSC と移動平均法の等価性を示した．

今後の課題としては，等価性の前提や問題の回避策をより厳密に示すこと，移動平均法に基づく説明とスペクトルクラスタリング [3] や伝達関数 [4] に基づく説明の関係を議論することが考えられる．

参考文献

[1] Keogh, E., Lin, J., and Truppel, W., Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research, IEEE Int'l Conf. on Data Mining ICDM-2003, pp.115–122 (2003).

[2] Chen, J. R., Making Subsequence Time Series Clustering Meaningful, IEEE Int'l Conf. on Data Mining ICDM-2005, pp.114–121 (2005).

[3] Ide, T., Why Does Subsequence Time-Series Clustering Produce Sine Waves?, Lecture Notes on Artificial Intelligence LNAI, pp.609–616 (2006).

[4] 藤巻遼平, 広瀬俊亮, 中田貴之, 部分時系列クラスタリングの周波数解析, 情報論的学習理論ワークショップ IBIS2007, pp.162–169 (2007).