k-Shape: Efficient and Accurate Clustering of Time Series

John Paoarrizos, Luis Gravano(Columbia University)
k-Shape: 効率的で正確な時系列クラスタリング

蒲田 涼馬 (Ryoma Gamada)
u455007@st.pu-toyama.ac.jp

富山県立大学大学院 工学研究科 電子・情報工学専攻
情報基盤工学講座

July 22, 2025

背景と目的

背景

時系列データ分析は金融，医療，工学分野で長年注目されてきた．
その中でもクラスタリングはデータから自動でパターンを発見できる強力な手法として広く利用されている．
しかし従来の時系列クラスタリング手法には計算コストの問題，またクラスタを代表する形状を平均化によってうまく捉えきれないという課題がある．

目的

時系列データをその形状に基づいてクラスタリングするための高精度かつ高効率的で汎用的な新しいアルゴリズムを提案する．

関連研究 1

動的時間伸縮法 (Dynamic Time Warping -DTW)

DTW は 2 つの時系列データ $x = (x_1, \ldots, x_m)$ と $y = (y_1, \ldots, y_m)$ があった際に時間的ずれや伸縮を許容しながらアライメントを行う手法
時間のずれや伸縮がある 2 つの波形でも，形状が最も似るように対応付けて類似度を計算する．

$$\gamma(i,j) = ED(i,j) + min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (1)$$

$ED(i,j)$ は 2 つの時系列の i 番目の点と j 番目の点の間の局所的な距離，min はそこまでの最短の距離を意味する．

k-means 法

与えられたデータ群を k 個のクラスタに分類するためのクラスタリング手法
利用者がクラスタの数 k を決め，クラスタの数だけ適当な点を設定
k-means では以下の目的関数 J が定義されており，J が小さいほど良いクラスタリングであると評価される (式 2 参照)
そして式 3 で重心が更新される

$$J = \sum_{k=1}^{K} \sum_{x_i \in S_k} ||x_i - \mu_k||^2 \tag{2}$$

$$\mu_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i \tag{3}$$

ここで K はクラスタの総数，S_k はクラスタ k に属するデータの集合，mu_k はクラスタ k の重心，$||x_i - \mu_k||^2$ はデータポイントと重心のユークリッド距離である．

提案手法 (k-Shape) について

k-Shape の距離尺度

k-Shape では距離尺度として Shape-Based Distance(SBD) を使う
SBD は 2 つの時系列データ x と y の形状がどれだけ似ているかを求めるために設計された距離尺度である.

$$SBD(x,y) = 1 - max_w(\frac{CC_w(x,y)}{\sqrt{R_0(x,x) \cdot R_0(y,y)}}) \tag{4}$$

$$CC_w(x,y) = R_{w-m}(x,y) \tag{5}$$

$$R_k(x,y) = \sum_{l=1}^{m-k} x_{l+k} \cdot y_l (k >= 0) \tag{6}$$

ここで m は時系列の長さ，k はシフト量，l は時系列内の位置，R は内積を意味する.

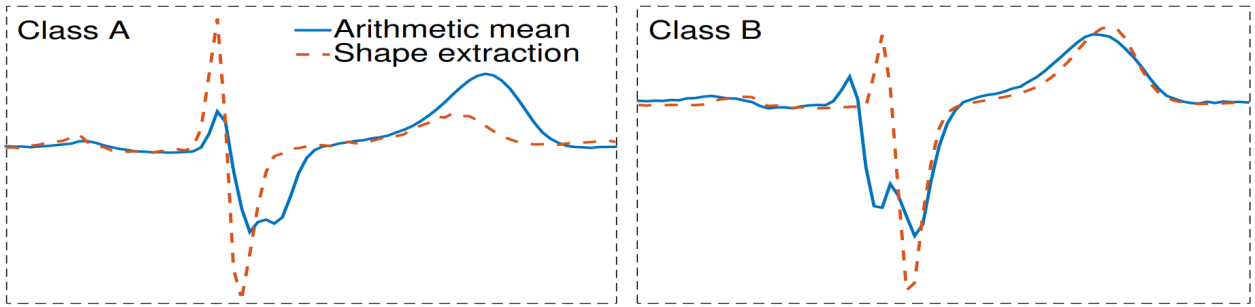
提案手法 (k-Shape) について

セントロイド計算とアルゴリズム

各クラスタを代表する形状 (セントロイド) は以下のように求められる.

$$\mu = \underset{z}{\operatorname{argmax}} \sum_{x \in S} Sim(z, x)^2 \tag{7}$$

ここで，μ は求めるセントロイド，S はクラスタに属する時系列データの集合，z はセントロイドの候補となるベクトル，$Sim(z, x)$ は 2 つの時系列の類似度を意味する.



k-Shape アルゴリズムの流れ

k-Shape アルゴリズムの流れ

- 初期化: クラスタ数 k を指定，入力された時系列データをランダムに k 個のクラスタに割り当てる．
- 反復計算: 更新ステップと割り当てステップを繰り返す．
  a. 更新ステップ: Shape Extraction を用いたセントロイドの計算
  b. 割り当てステップ: すべての時系列データについて，k 個の新しいセントロイドとの距離 SBD を用いて計算，割り当てを行う．
- 終了: クラスタのメンバー割り当てに変化がなくなったら最終的なクラスタリング結果と各クラスタのセントロイドを出力する．

実験の概要

実験の概要

提案手法が既存の手法と比較して精度と効率性において優れているかを実証する

- データセット: 48 種類の公開されている時系列データセットを使用. 合成データと実データの両方が含まれ，様々な分野を網羅.
- 評価指標: Rand Index を使用してクラスタリング結果の正確さを評価

$$RandIndex = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

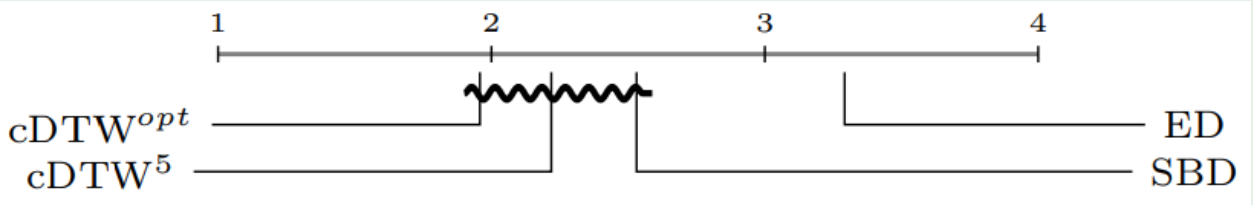
また，フリードマン検定およびネメニー検定を行い，手法の性能が統計的に差があるのかを検定，手法間の性能にどれほどの有意的な差があるのかを検定する．効率性: CPU 実行時間を計測.

- 比較対象: K-measn + ユークリッド距離法，K-means + DTW(k-DBA)，KSC，K-medoids，階層的クラスタリング，スペクトラムクラスタリング

実験結果

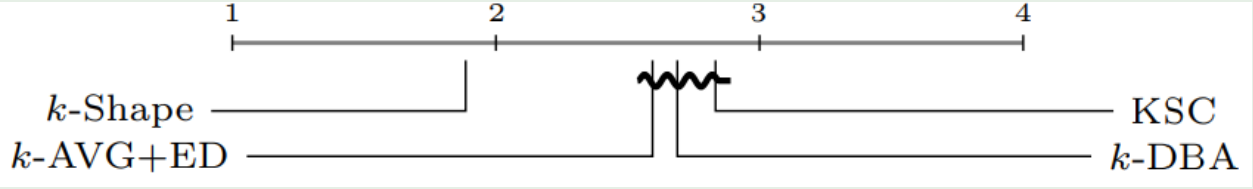
距離尺度 SBD の評価

距離尺度 SBD はユークリッド距離を優位に上回り，最先端の cDTW と同等の精度であった
計算速度は cDTW より約 10 倍速くなっていた
また，cDTW はクラスタ数以外のハイパーパラメータも存在するため，そこでも差別化可



k-Shape の評価

k-Shape は比較対象となったほかの手法の中で有意的に最高の性能
当時最先端の手法と比較して 100 倍高速
ユークリッド距離よりは遅いが，精度の大幅な向上を考えると有用である



おわりに

まとめ

この論文では SBD と Shape Extraction を用いた k-Shpae を提案し，その精度と計算速度を検証した
結果として，高精度でかつ計算速度が速いということがわかった

この論文を選んだ理由

この論文を選んだ理由

金融時系列の分析を行ううえで非定常性の考慮は多くの論文で課題として挙げられている.
修士研究では時系列を考慮すると同時に非定常性も考慮した分析をしたい.
また，クラスタリングを行うことで可読性が向上するかもしれない.

今後の方針

- E2E シンボリック回帰の学習部分のプログラム
- 学習時間と精度の確認
  学習時間が長すぎる．精度がイマイチ，いつまでたっても実装できない
  →学習済みモデルからファインチューニングで時系列，非定常性を考慮させていく
  実装ができてすべてがうまくいきそうなら
  →学習プログラム自体を書き換えていく.
  同時並行でクラスタリングによる非定常性を考慮した分析をやっていく.