

Matrix Profile XIII: Time Series Snippets (A New Primitive for Time Series Data Mining)

1 概要

1.1 どんな論文?

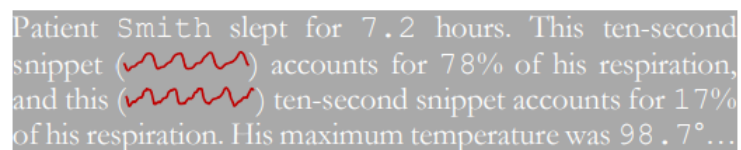
Matrix Profile を用いて時系列データから頻繁に表れるような部分時系列 (snippets) を抽出

1.2 似ている手法

モチーフ検出、部分時系列クラスタリング、shapelets など

1.3 メリット

- 長い時系列から代表的な (頻繁に表れる) 部分時系列を抽出できる
- 抽出した snippets がどの程度支配的かの割合 (論文では coverage に対応) を得られる
論文中の例:



Patient Smith slept for 7.2 hours. This ten-second snippet (~~~~~) accounts for 78% of his respiration, and this (~~~~~) ten-second snippet accounts for 17% of his respiration. His maximum temperature was 98.7°...

図1 train データの例

- ノイズに対してロバスト (MPdist を使っているから?)

2 手法

前提知識: matrix profile, MPdist

snippets の長さを m に設定する。また、扱う時系列の長さを n とする。

- snippets の探索 (貪欲法)

Q = 値が全て Inf のベクトル (長さ $m - n - 1$)

$C = \emptyset$ (snippets を格納するリスト)

1. 時系列を n/m 等分する (等分された部分時系列を左から $T_1, T_2, \dots, T_{n/m}$ とする)

2. $T_1, T_2, \dots, T_{n/m}$ のそれぞれについて MPdist ベクトルを作成する

3. for num=1:k

– 各々の MPdist ベクトル D_i について ProfileArea を求める

ProfileArea _{i} = sum(min(D_i, Q))

– $j = \arg \min_i \text{ProfileArea}_i$ を計算し、 C_{num} に T_j を格納

– $Q = \min(D_i, Q)$ に更新

- coverage の求め方

– T_i の coverage は、 T_i と Q で同じ値を取る時間点の割合

3 感想・メモ

- アルゴリズムはものすごく単純

- 貪欲法の部分はもう少し最適化できそうかも (精度的に)

- 得られたすべての snippets の coverage を足すと必ず 100%(以上) になる

- 計算オーダーが $O(n^2 \times (n - m)/m)$

($O(n^2)$: それぞれの部分時系列に対しての MPdist ベクトルの計算量、 $O((n - m)/m)$: 等分された部分時系列の数)

らしいが、等分された部分時系列の長さは m なので、 $O(nm \times (n - m)/m) = O(n \times (n - m))$ のような気がする・・・

($n \gg m$ の時、前者と後者のオーダーがそれぞれ $O(n^3)$ と $O(n^2)$ となり大きな差があるので、どちらが合っているのか要調査)