Sequence Classification(序列分类)

Real value sequence as exemplified

刘精昌

April 27, 2016

Table of Contents

sequence

what is sequence?

- DNA and protein sequence
- The time series of heart rates
- Trend of stock

How to represent sequence?

- An ordered list of the symbols, such as ACCCCCGT
- A sequence of real values, such as 0.1,0.3,0.5,0.1,...

sequence classification

Applications of sequence classification

- gait analysis
- speech recognition
- learn the functions of a new protein

sequence classification

Applications of sequence classification

- gait analysis
- speech recognition
- learn the functions of a new protein

task

A sequence may carry a class label. Given L as a set class labels, the task of $(conventional)$ *sequence classification* is to learn a *sequence classifier* C, which is a function mapping a sequence s to a class label $l \in L$,written as,

$$C : s \rightarrow l, l \in L$$

methods and problems

methods

1NN, 待分类序列的 label 即距离其最近的序列的 label

problems

How to measure the distance between two sequence?

Table of Contents

Table of Contents

Euclid distance

The simplest distance is Euclid distance:

$$dist\left(s, s^{'}\right) = \sqrt{\sum_{i=1}^{L} \left(s\left[i\right] - s^{'}\left[i\right]\right)^{2}}$$

And other similar distance.

Weakness of Euclid distance

- 需要满足两序列长度相同。
- 设想这样一种情况。在步态分析中，同一测试者的步速可能不同，或者在某时间段上存在着加速和减速。那么对于其两段步态序列，比较相似的步态之间可能会有一定的时间差，而上面的这些距离测度只会将同一时刻的步态相比较。也就是说，上面的这些距离测度不能反映出序列比较中的错位。

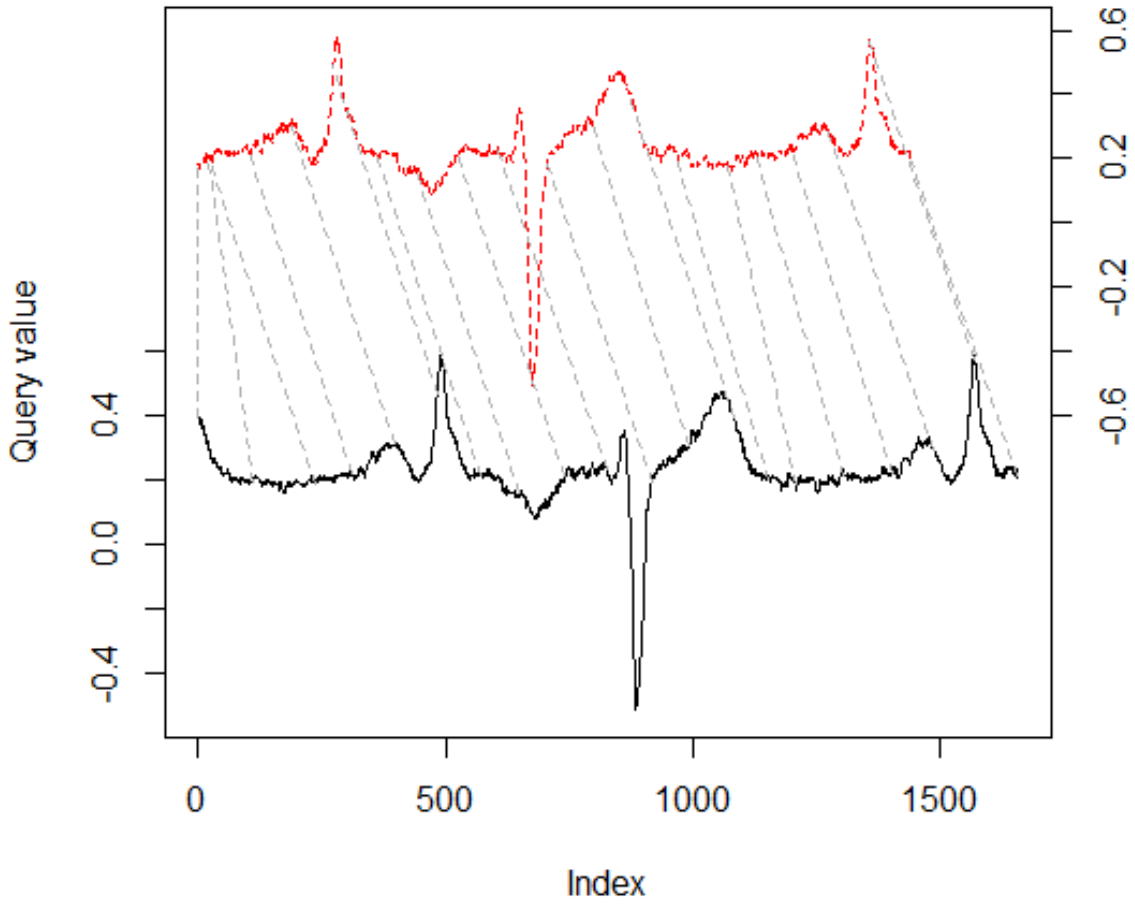


Figure: DTW 示意图

Table of Contents

Definitions

- $Q = q_1, q_2, \cdots, q_i, \cdots, q_m, C = c_1, c_2, \cdots, c_j, \cdots, c_n$
- $D(i^{th}, j^{th}) = d(q_i, c_j) = (q_i - c_j)^2$
- warping path:

$$W = w_1, w_2, \cdots, w_k, \cdots, w_K \quad max(m,n) \leq K \leq m + n - 1$$

$$w_k = (i,j)_k$$

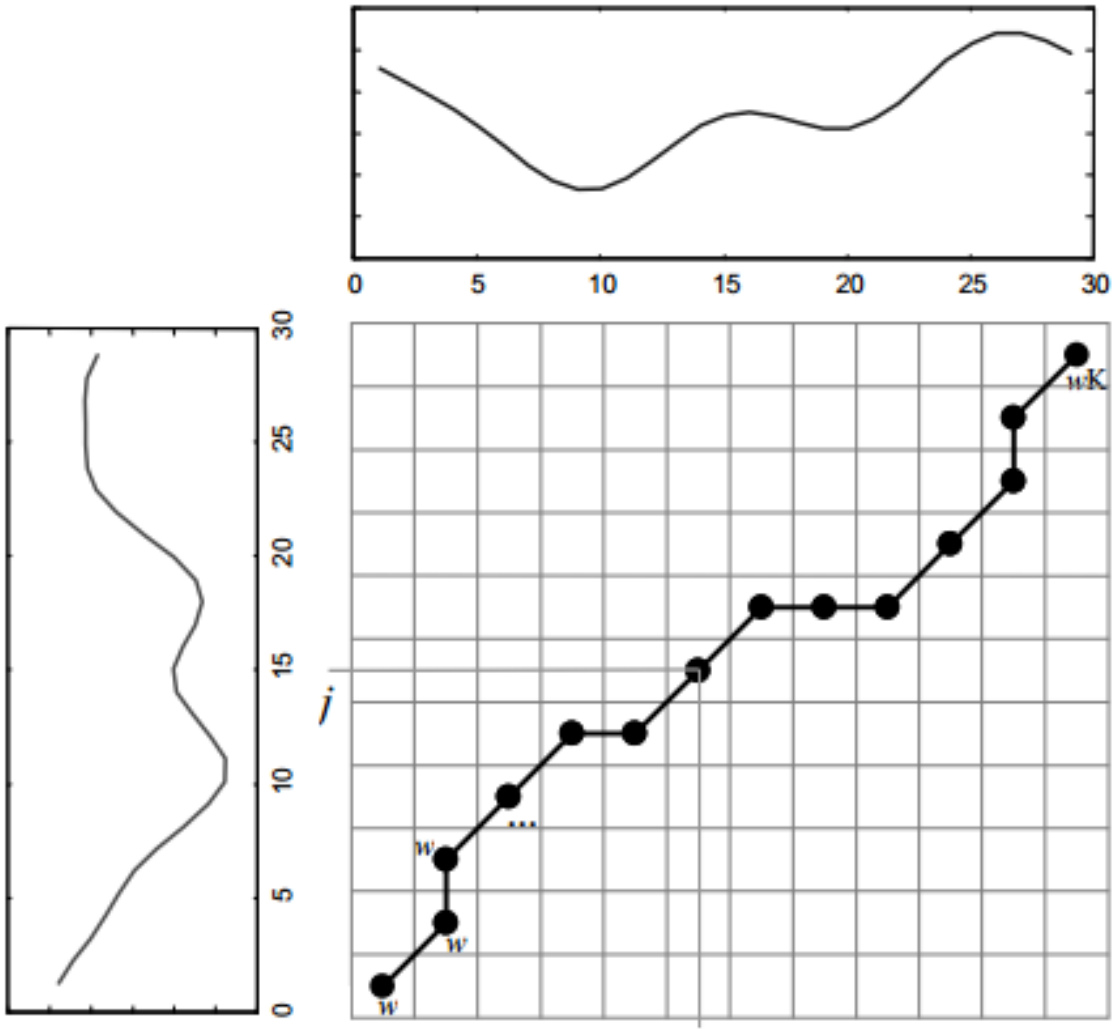


Figure: warping path 示意图

Constraints

- **Boundary conditions:** $w_1 = (1,1)$ and $w_K = (m,n)$
- **Continuity:** Given $w_k = (a,b)$ then $w_{k-1} = (a^{'},b^{'})$ where $a - a^{'} \leq 1$ and $b - b^{'} \leq 1$
- **Monotonicity:** Given $w_k = (a,b)$ then $w_{k-1} = (a^{'},b^{'})$ where $a - a^{'} \geq 0$ and $b - b^{'} \geq 0$

target and evaluation

Target

minimizes the warping cost:

$$DTW(W) = \sum_{k=1}^{K} d\left(w_{ki}, w_{kj}\right)$$

$d\left(w_{ki}, w_{kj}\right)$: the distance between two data point indexes(one from Q and one from C) in the k^{th} element of the warp path.

Evaluation

$\gamma(i,j)$: cumulative distance

$$\gamma(i,j) = d\left(q_i, c_j\right) + \min\left(\gamma(i-1,j), \gamma(i,j-1), \gamma(i-1,j-1)\right)$$

Algorithm 1 Calculate DTW

Require: $s : array[1..m], t : array[1..n]$
Ensure: $DTW[m,n]$

1. $DTW := [0..m, 0..n]$
2. **for** $i := 0$ to m **do**
3. $\quad DTW[i,0] := inf$
4. **end for**
5. **for** $j := 0$ to n **do**
6. $\quad DTW[0,j] := inf$
7. **end for**
8. $DTW[0,0] := 0$
9.
10. **for** $i := 1$ to m **do**
11. $\quad$ **for** $j := 1$ to n **do**
12. $\quad\quad cost := d(s[i], t[j])$
13. $\quad\quad DTW[i,j] := cost + min(DTW[i-1,j], DTW[i,j-1], DTW[i-1,j-1])$
14. $\quad$ **end for**
15. **end for**

Trace back the best path

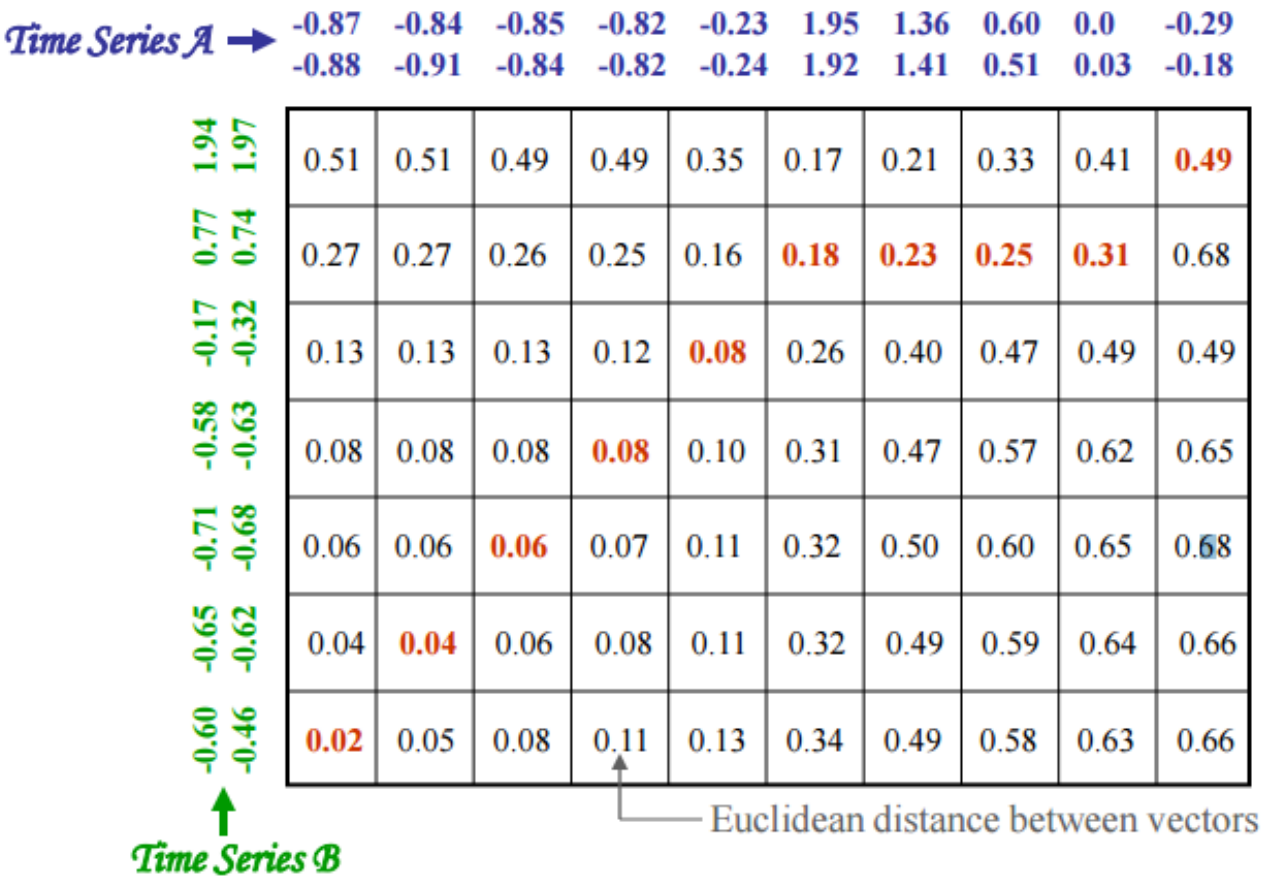


Figure: A greedy search is performed that evaluates cells to the left,down,and diagonally to the bottom-left

Table of Contents

warp window

- An obvious observation is that an intuitive alignment path is unlikely to drift/very far from the diagonal

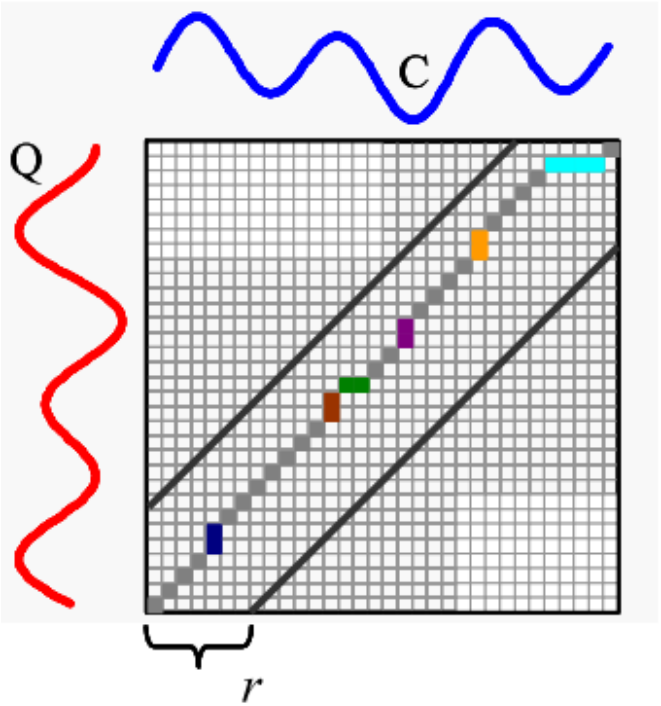


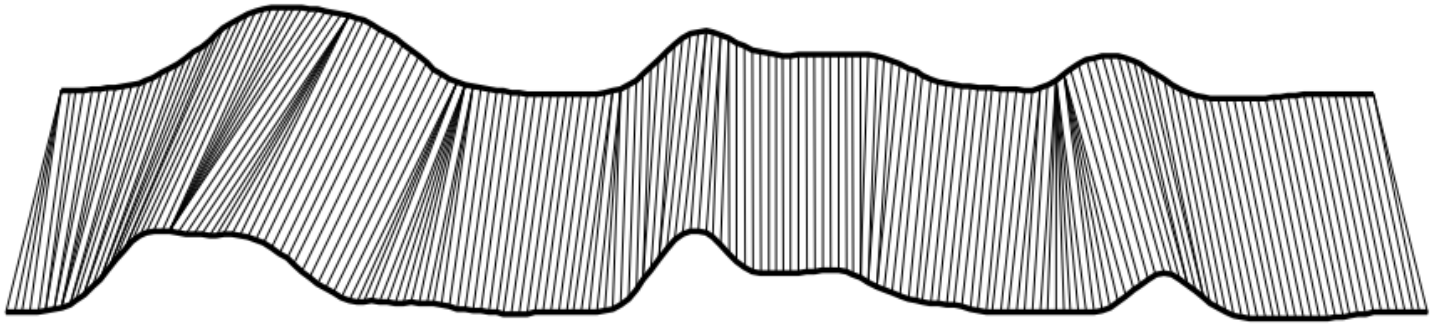
Figure: warp window

Algorithm 2 Calculate DTW with warp window

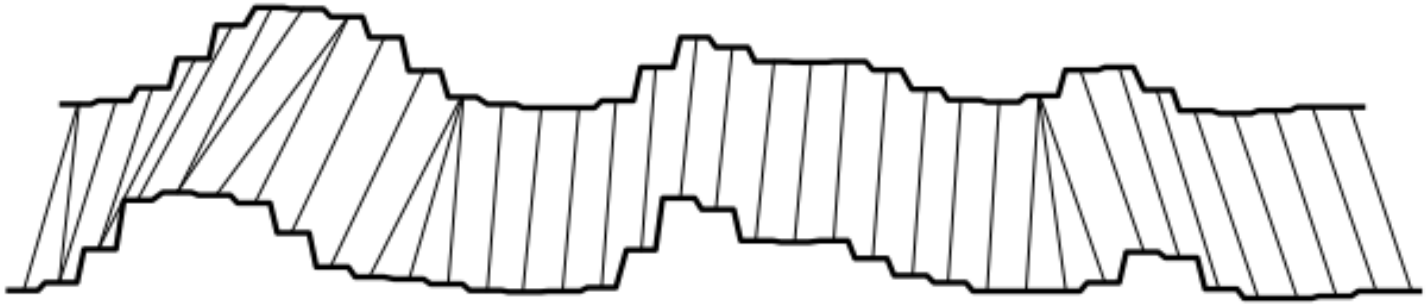
Require: $s : array[1..n], t : array[1..m], w$:warp window
Ensure: $DTW[n,m]$

1. $DTW := array[0..n, 0..m]$
2. $w := max(w, |n-m|)$
3. **for** $i := 0$ to n **do**
4. **for** $j := 0$ to m **do**
5. $DTW[i,j] := inf$
6. **end for**
7. **end for**
8. $DTW[0,0] := 0$
9.
10. **for** $i := 1$ to n **do**
11. **for** $j := max(1, i-w)$ to $min(m, i+w)$ **do**
12. $cost := d(s[i], t[j])$
13. $DTW[i,j] := cost + min(DTW[i-1,j], DTW[i,j-1], DTW[i-1,j-1])$
14. **end for**
15. **end for**

Piecewise aggregate representation



(a) 原始 DTW 对齐



(b) PAR 处理后的 DTW 对齐

Figure: PAR 处理示意图

FastDTW

Time and Space Complexity

- **DTW:**$O(n^2)$
- **FastDTW:**$O(n)$

Three key operations

1. Coarsening(粗化)
2. Projection（投影）
3. Refinement

FastDTW

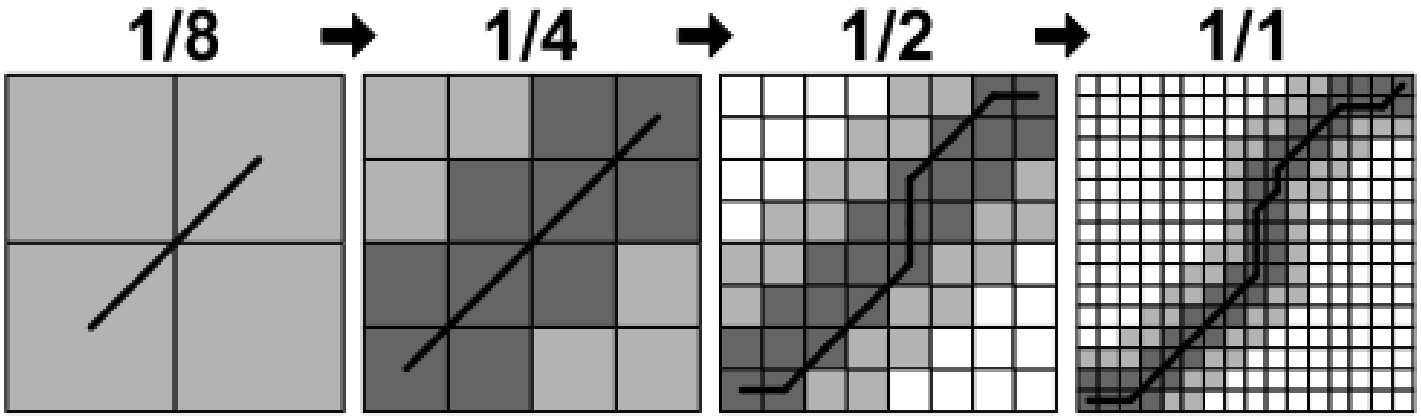


Figure: FastDTW 示意图

Time Complexity of FastDTW

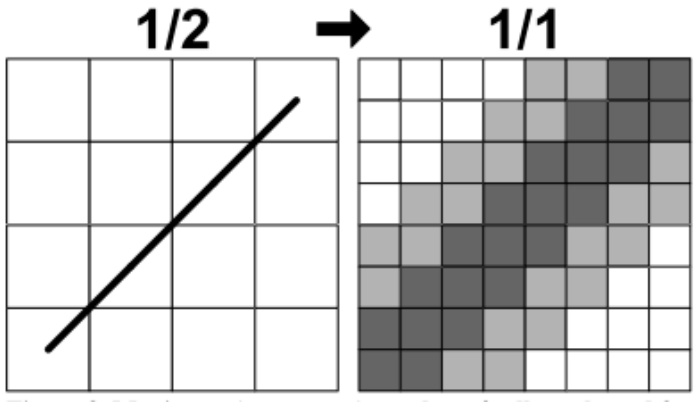


Figure: Maximum number of cells evaluated for a radius of 1

maximum number of cells : $3N + 2 * (2Nr) = N(4r + 3)$

Total number of cells filled:

$$N(4r+3) + \frac{N}{2}(4r+3) + \frac{N}{2^2}(4r+3) + \cdots = 2N(4r+3)$$

Time Complexity of FastDTW

Time Complexity

1. number of cells calculated: $2N(4r + 3)$
2. creat the coarser resolutions: $4N$
3. determining the warp path by tracing throungth the matrix: $4N$

Total FastDTW time complexity

$$N(8r + 14)$$

Space Complexity of FastDTW

Space Complexity

1. Space of resolutions: $4N$
2. Space of distance matrix: $N(4r + 3)$
3. Space complexity of storing the warp path: $4N$

Total FastDTW space complexity

$$N(4r + 11)$$

Table of Contents

experimental result

data: 在一段间隔上加上随机误差而产生

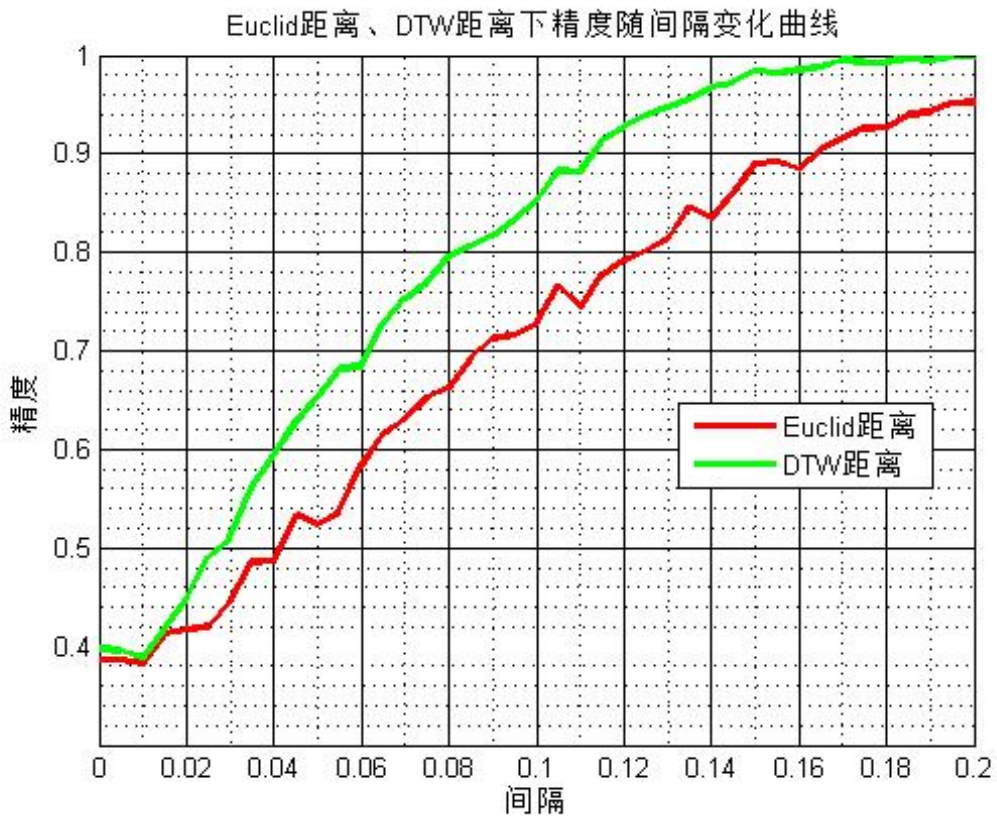


Figure: Euclid、DTW 距离下分类精确度随间隔变化曲线

experimental result

TSDMA数据集

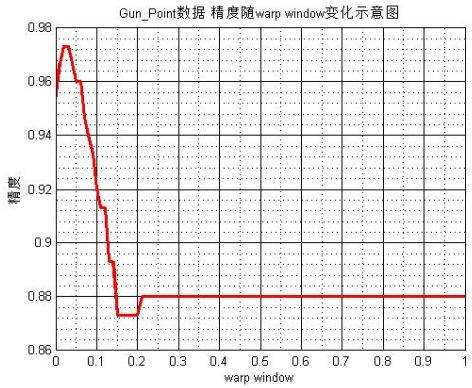
name	Computers	Trace	FaceFour	WordsSynonyms
lasses	2	4	4	25
training set size	250	100	24	267
test set size	250	100	88	638
sequence length	720	275	350	270
error rate (Euclid)	0.424	0.24	0.21591	0.38245
error rate (DTW)	0.332	0.01	0.15909	0.32445

experimental result

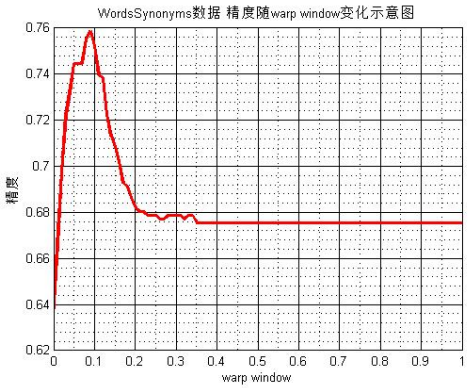
TSDMA数据集

name	Gun_Point	Plane	StrawBerry
lasses	2	7	2
training set size	50	105	370
test set size	150	105	613
sequence length	150	144	235
error rate (Euclid)	0.086667	0.038095	0.06199
error rate (DTW)	0.12	0	0.066884

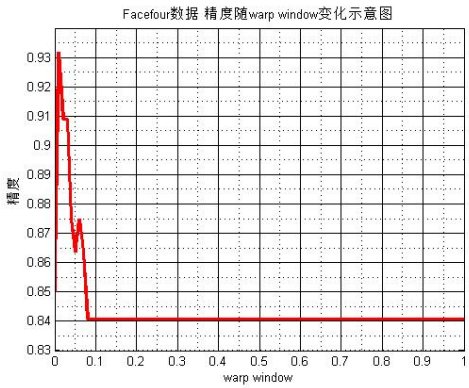
experimental result



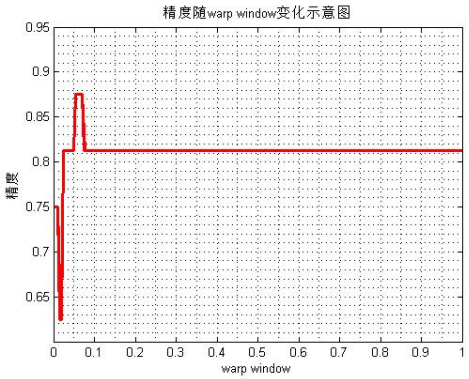
(a) gun point



(b) WordsSunonyms



(c) facefour



(d) interval:0.135

Figure: 精度随 warp window 变化示意图

Table of Contents

Table of Contents

Introdiction

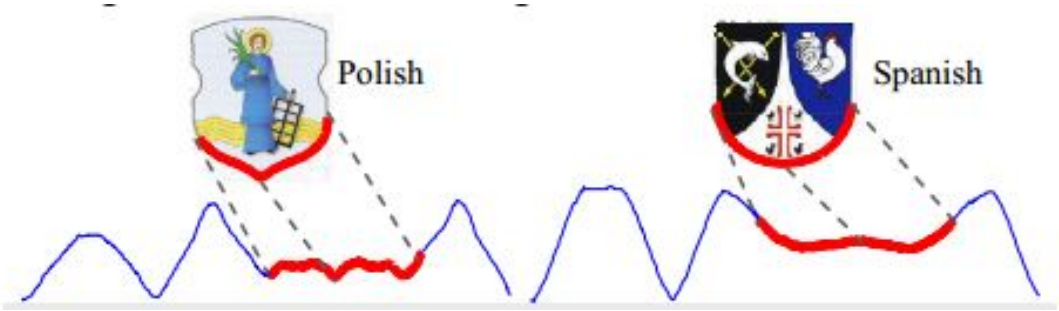


Figure: shapelets are subsequences which are in some sense maximally representative of a class

Advantages

1. Provide interpretable results
2. More accuracy/robust on some datasets
3. Faster,$O(ml)$, m :length of query sequence, l :length of shapelets

Definition

$$SubsequenceDist(T,S) = min(Dist(S',S)), \textbf{ for } S' \in S_T^{|S|}$$

Optimal Split Point(OSP) . A sequence dataset **D** consists of two classes,A and B.For a shapelet candidate S,we choose some distance threshold d_{th} and split **D** into D_1 and D_2,such that for every time series object $T_{1,j}$ in D_1,$SubsequenceDist(T_{1,i},S) < d_{th}$ and for every time series object $T_{2,i}$ in D_2,$SubsequenceDist(T_{2,i},S) > d_{th}$. An Optimal Split Point is a distance threshold that

$$Gain\left(S, d_{OSP(D,S)}\right) \geq Gain\left(S, d_{th}^{'}\right)$$

for any other distance threshold $d_{th}^{'}$.

Definition

Shapelet. Given a time series dataset **D** which consists of two classes, A and B,*shapeletD* is a subsequence that, with its corresponding optimal split point,

$$Gain\left(shapelet\left(D\right), d_{OSP(D,shapelet(D))}\right) \geq Gain\left(S, d_{OSP(D,S)}\right)$$

for any other subsequence S.

Table of Contents

Algorithm 3 Brute force algorithm for finding shapelet

Require: dataset *D*,*MAXLEN*,*MINLEN*
Ensure: bsf_shapelet

```
1. candidates := GenerateCandidates(D,MAXLEN,MINLEN)
2. bsf_gain := 0
3. for S in candidates do
4.     gain:=CheckCandidate(D,S)
5.     if gain>bsf_gain then
6.         bsf_gain :=gain
7.         bsf_shapelet :=S
8.     end if
9. end for
```

Speedup methods

Subsequence Distance Early Abandon

- Stop distance calculations once the partial distance exceeds the minimum distance known so far.

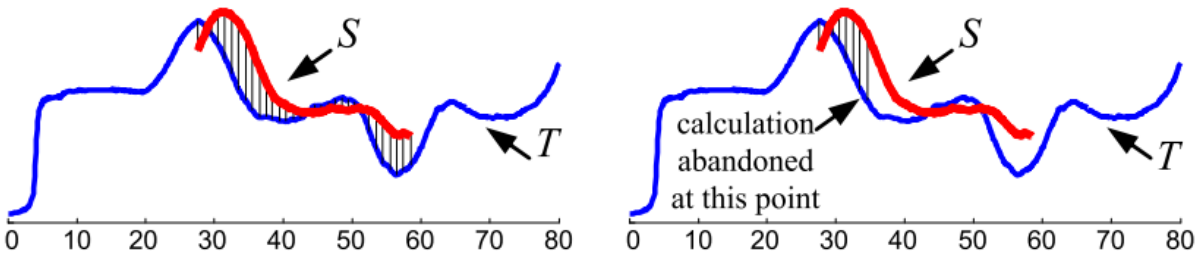
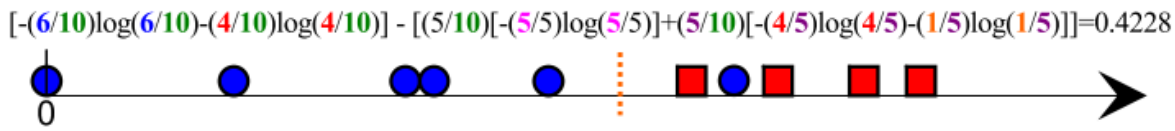
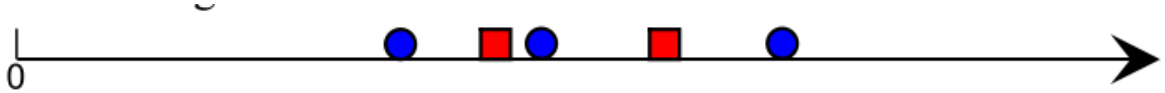


Figure: Subsequence Distance Early Abandon

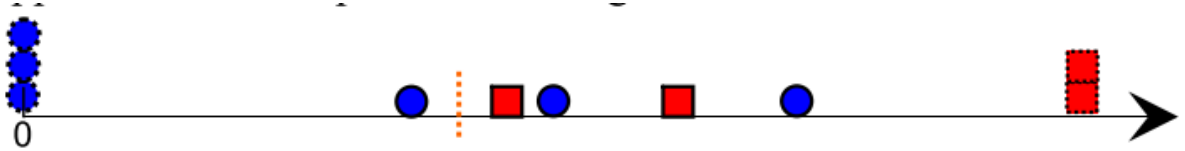
Admissible Entropy Pruning



(a) 排序依照数据集序列到候选序列距离，并计算当前的信息增益



(b) 已计算了数据集中的五个序列到候选序列的距离



(c) 数据集剩余序列使得信息增益最大的极端位置

Figure: Entropy Prunning 示意图

Q & A

谢谢观看