Fast Global Alignment Kernels

Time Series

05.05.2015
Rodrigo Frassetto Nogueira
Thanos Papadopoulos
Agenda

- Motivation
- Dynamic Time Warping (DTW)
- Global Alignment (GA) Kernels
- Experiments
- Conclusion
Agenda

• Motivation
  • Time series Introduction
  • Problem Formulation
    • Dynamic Time Warping (DTW)
    • Global Alignment (GA) Kernels
    • Experiments
    • Conclusion
Example #1

• Stock Price Forecasting

Example #2

- Caltrans Performance Measurement System (PeMS)
  - occupancy rate in San Francisco bay area freeways - [0, 1]
  - 963 sensors (different car lanes)
  - data every 10'
  - time series per day: dimension - 963, length - 6*24=144
Example #3

- Speech signals
Agenda

• Motivation
  • Time series Introduction

• Problem Formulation
  • Dynamic Time Warping (DTW)
  • Global Alignment (GA) Kernels
  • Experiments
  • Conclusion
Problem Formulation

• Goal: Find adequate kernels for time series
  • handle variable length
  • be positive definite
  • low computational cost
How we start?

Similarity Measure: \( K(x, y) = \langle \phi(x), \phi(y) \rangle \)
Agenda

- Motivation

- **Dynamic Time Warping (DTW)**
  - Global Alignment (GA) Kernels
  - Experiments
  - Conclusion
Dynamic Time Warping

- pairwise comparisons?

- alignment:
  associate each element of sequence $X$ to one or more elements of sequence $Y$ and vice-versa
DTW Cost Matrix
Optimal Alignment
DTW definition

- warping functions: $\pi_1(i), \pi_2(j)$
DTW definition

- warping functions: \( \pi_1(i), \pi_2(j) \)
- moves: \( \rightarrow, \uparrow, \nearrow \)
DTW definition

• warping functions: $\pi_1(i), \pi_2(j)$

• moves: $\rightarrow, \uparrow, \nearrow$

• cost per alignment $\pi$:

$$D_{x,y}(\pi) = \sum_{i=1}^{\mid\pi\mid} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)})$$

• optimal alignment:

$$DTW(x, y) = \min_{\pi \in A(n,m)} D_{x,y}(\pi)$$
Why not DTW?

- not PDS
- high computational cost, $O(dnm)$
Agenda

- Motivation
- Dynamic Time Warping (DTW)
- **Global Alignment (GA) Kernels**
  - **Definition**
  - Diagonal Dominance
  - Positive Definiteness
  - Computational Cost
- Experiments
- Conclusion
GA kernels

- soft maximum motivation, \( \log(\sum_i e^{x_i}) : \)

\[
k_{GA} = \sum_{\pi \in A(n,m)} e^{-D_{x,y}(\pi)} = \sum_{\pi \in A(n,m)} e^{-\sum_{i=1}^{||\pi||} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)})}
\]

- by defining \( \kappa = e^{-\varphi} : \)

\[
k_{GA} = \sum_{\pi \in A(n,m)} \prod_{i=1}^{||\pi||} \kappa(x_{\pi_1(i)}, y_{\pi_2(i)})
\]

- whole spectrum of costs/alignments
Agenda

- Motivation
- Dynamic Time Warping (DTW)
- **Global Alignment (GA) Kernels**
  - Definition
  - **Diagonal Dominance**
  - Positive Definiteness
  - Computational Cost
- Experiments
- Conclusion
Diagonal Dominance

• off-diagonal entries far smaller than trace (Gram matrix)

• Solution:

\[ \kappa = e^{-\lambda \varphi(x_{\pi_1(i)}, y_{\pi_2(i)})}, \quad \lambda > 0 \]
Behavior in the limits of $\lambda$

Let $\{X_1, X_2, ..., X_p\}$ be a sample of time series

$$k_{GA}(x, y) = \sum_{\pi \in A(n,m)} \prod_{i=1}^{\left|\pi\right|} e^{-\lambda \cdot \varphi(x_{\pi_1(i)}, y_{\pi_2(i)})}$$

1. All samples have same length $n$

- When $\lambda \to \infty$, $k_{GA} \to I_p$
- When $\lambda \to 0$, $k_{GA} \to D(n, n) \cdot 1_{p,p}$
Delannoy numbers

\[ D(n, m) = D(n, m - 1) + D(n - 1, m) + D(n - 1, m - 1) \]
Behavior in the limits of $\lambda$

- Let $\{X_1, X_2, ..., X_p\}$ be a sample of time series

$$k_{GA}(x, y) = \sum_{\pi \in A(n, m)} \prod_{i=1}^{\vert \pi \vert} e^{-\lambda \cdot \varphi(x_{\pi_1(i)}, y_{\pi_2(i)})}$$

2. Samples have different lengths

- When $\lambda \to \infty$, $k_{GA} \to I_p$

- When $\lambda \to 0$, $k_{GA}(X_i, X_j) \to D(|X_i|, |X_j|)$
Diagonal Dominance

• problem arises when length varies and $\lambda$ tends to 0

• $n$ sequences with length 1 to $n$

• Empirically: for $\lambda \approx 0 \Rightarrow \frac{1}{2} \leq \frac{n}{m} \leq 2$
Agenda

• Motivation
• Dynamic Time Warping (DTW)

• **Global Alignment (GA) Kernels**
  • Definition
  • Diagonal Dominance
  • **Positive Definiteness**
  • Computational Cost

• Experiments
• Conclusion
Positive Definiteness

- if $\kappa$ is p.d. what about $k_{GA}$?

- mapping kernels: $k(x, y) = \sum_{(x_i, y_i) \in M(x, y)} \kappa_l(x_i, y_i)$

$\kappa$ is a local kernel on substructures of $x, y$

$M$ is a mapping set
GA kernels as Mapping kernels

• if $\kappa$ is p.d. what about $k_{GA}$?

• For GA kernels: $\kappa_l(x_{\pi_1}, y_{\pi_2}) = \prod_{i=1}^{\lvert \pi \rvert} \kappa(x_{\pi_1(i)}, y_{\pi_2(i)})$

$M_{GA}(x, y) = \{(x_{\pi_1}, y_{\pi_2}) \mid \pi = (\pi_1, \pi_2) \in A(n, m)\}$

• Theorem 1: $k_{GA}$ is p.d. if and only if M is transitive

$(x_i, y_i) \in M(x, y), (y_i, z_i) \in M(y, z) \Rightarrow (x_i, z_i) \in M(x, z)$

• Lemma 1: $M_{GA}$ is not transitive
Constraints on $\kappa$

- **Theorem 2**: If $\kappa$ is p.d. and $\frac{\kappa}{1 + \kappa}$ is p.d., $k_{GA}$ is p.d.

- **Geometric Divisibility (g.d.)**:
  
  - mapping $\tau(x) = \frac{x}{1 + x}$ : $R_+ \rightarrow [0, 1)$
  
  - inverse mapping $\tau^{-1}(x) = \frac{x}{1 - x}$ : $[0, 1) \rightarrow R_+$

  - $f$ is g.d. if $\tau f$ is p.d.

- **Lemma 2**: If $\kappa$ is g.d. then $k_{GA}$ is p.d.
Constraints on \( \kappa \)

- Infinite divisibility (i.d.): \( \kappa \) is i.d. iff \(-\log(\kappa)\) is n.d.

- Lemma 3: For any i.d. kernel \( \kappa \) s.t. \( 0 < \kappa < 1 \), \( \tau^{-1} \kappa \) is g.d. and i.d.

- how to construct a local kernel?

- Example: Gaussian kernel, \( \kappa_\sigma \), is i.d., thus \( \tau^{-1}(\frac{\kappa_\sigma}{2}) \) is i.d. and g.d.

Also, \( \varphi = -\log(\tau^{-1}(\frac{\kappa_\sigma}{2})) \) is n.d. and we set \( \kappa = e^{-\lambda \varphi} \)
Agenda

- Motivation
- Dynamic Time Warping (DTW)
- **Global Alignment (GA) Kernels**
  - Definition
  - Diagonal Dominance
  - Positive Definiteness
  - **Computational Cost**
  - Experiments
  - Conclusion
Time Complexity

- DTW time complexity: O(dnm)

What to do?

- Ignore “bad” alignments - keep alignments close to the diagonal:

\[
D_{x,y}^\gamma (\pi) = \sum_{i=1}^{\left|\pi\right|} \gamma_{\pi_1(i),\pi_2(i)} \cdot \varphi(x_{\pi_1(i)}, y_{\pi_2(i)})
\]

\[
\gamma_{i,j} = \begin{cases} 
1, & |i-j| < T \\
\infty, & |i-j| \geq T 
\end{cases}
\]
\[ M_{i,j} = \kappa(x_i, y_j)(M_{i-1,j-1} + M_{i,j-1} + M_{i-1,j}) \]

\[(2T - 1) \cdot \min(n, m) - T(T - 1)/2 \Rightarrow O(2T\min(n, m))\]
Triangular GA Kernels

- if $\kappa$ is infinite divisible, and $\omega$ is a p.d. kernel in $\mathbb{N}$, then $\tau^{-1}(\omega \otimes \kappa)$ as a local kernel gives a p.d. GA kernel

- common choice for $\omega$:

$$\omega(i, j) = \left(1 - \frac{|i - j|}{T}\right)_+$$

- Example: $\tau^{-1}(\omega \otimes \frac{1}{2}\kappa_\sigma)(i, x; j, y) = \frac{\omega(i, j)\kappa(x, y)}{2 - \omega(i, j)\kappa(x, y)}$
Agenda

• Motivation
• Dynamic Time Warping (DTW)
• Global Alignment (GA) Kernels

• Experiments

• Conclusion
Kernels to compare

- **DTW:** \( k_{DTW} = e^{-\frac{1}{t}DTW} \)

- **DTW SC:** \( k_{SC} = e^{-\frac{1}{t}DTW_{SC}} \)

\[
DTW_{SC}(x, y) = \min_{\pi \in A(n,m)} D^\gamma_{x,y}(\pi)
\]

- **DTAK:** \( (k_{DTAK}(x, y))^{\frac{1}{t}} \)

\[
k_{DTAK}(x, y) = \max_{\pi \in A(n,m)} \sum_{i=1}^{|\pi|} \kappa_\sigma(x_{\pi_1(i)}, y_{\pi_2(i)})
\]

- **GA:** \( \kappa = e^{-log\left(\tau^{-1}\left(\frac{1}{2}\kappa_\sigma\right)\right)} \)

- **TGA:** \( \kappa = \tau^{-1}(\omega \otimes \frac{1}{2}\kappa_\sigma) \)
Datasets

<table>
<thead>
<tr>
<th>Database</th>
<th>$d$</th>
<th>$n$ range, med($n$)</th>
<th>classes</th>
<th># points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese Vowels</td>
<td>12</td>
<td>7-29, 15</td>
<td>9</td>
<td>640</td>
</tr>
<tr>
<td>Libras</td>
<td>2</td>
<td>45</td>
<td>15</td>
<td>945</td>
</tr>
<tr>
<td>Handwritten Characters</td>
<td>3</td>
<td>60-182, 122</td>
<td>20</td>
<td>2858</td>
</tr>
<tr>
<td>AUSLAN</td>
<td>22</td>
<td>45-136, 55</td>
<td>95</td>
<td>2465</td>
</tr>
<tr>
<td>PEMS</td>
<td>963</td>
<td>144</td>
<td>7</td>
<td>440</td>
</tr>
</tbody>
</table>
Classification error rates

Mean/StD of test error rates (%)

AUSLAN  LIBRAS  PEMS

DTW  DTW SC  DTAK  GA  TGA

JV  HW
Performance and speed
Agenda

• Motivation
• Dynamic Time Warping (DTW)
• Global Alignment (GA) Kernels
• Experiments

• **Conclusion**
Conclusion

$O(T \cdot \min(n, m))$

Fast Global Alignment Kernels

wide spectrum of alignments

DTW based

PDS variable size