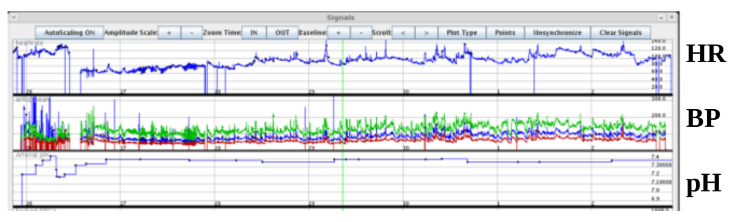
Deep ExpeCted Alignment DistancE (DECADE): A Deep Metric Learning Model for Multivariate Time Series

Zhengping Che, Xinran He, Ke Xu, Yan Liu
University of Southern California
{zche, xinranhe, xuk, yanliu.cs}@usc.edu

USC UNIVERSITY OF SOUTHERN CALIFORNIA

Background

- **Determining similarity (or distance)** between multivariate time series is useful and fundamental



  - Find similar patients for better diagnosis and decision making
  - Verify whether two voice clips are from the same speaker
- Finding a good multivariate time series similarity is extremely challenging
  - Complex *temporal* dependencies
  - *Variable* lengths of time series
- *No universal* similarity measure works best across all time series applications
  - Learning a *data-dependent* distance metric is vital

Motivation and Comparisons

- Three desired properties of good time series similarity measures
  - What kind of **local** distance to use
    - × Predefined local distance
    - ✓ Flexible data-dependent local distance for multivariate data
  - Whether to **align** the time series
    - × Do not take the (pairwise temporal) alignment
    - ✓ Use alignment to capture temporal dependencies
  - Whether to have a **valid distance metric** which satisfies triangle inequality
    - × Not a valid metric/pseudo-metric
    - ✓ A valid metric which can be used for e.g., kernel methods, and fast nearest neighbor search
- Comparison of some common time series similarities and our proposed model

		Data-dependent local metric	Considering alignment	Valid metric
MDTW	[Berndt, James, 1994]	No	Single	No
GAK	[Cuturi et al., 2007]	No	Multiple	*Yes*[1]
MSA	[Hogeweg, Ben, 1984]	No	Single	Yes
ML-TSA	[Garreau et al., 2014]	Yes (Linear)	*Single*[2]	No
LDMLT-TS	[Mei et al., 2016]	Yes (Linear)	Single	No
MaLSTM	[Mueller, Aditya, 2016]	Yes (Deep)	No	Yes
DECADE	Proposed in this work	Yes (Deep)	Multiple	Yes

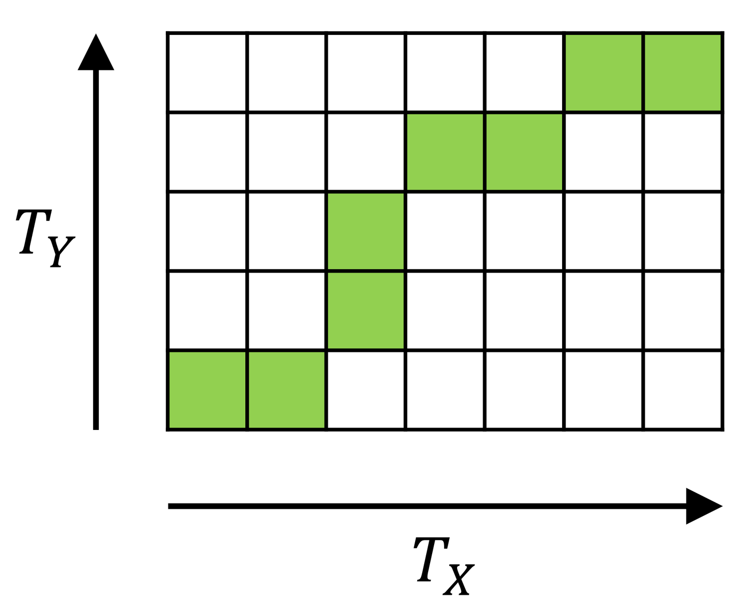
[1] *Constraints on local kernel selection;* [2] *Ground-truth alignment is required for training.*

Alignment and Distance on an Alignment

- $\boldsymbol{X} \in \mathbb{R}^{p \times T_X}$: a time series with p features and T_X time steps
- An *alignment* $\boldsymbol{A}$ of two time series $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be defined as a pair of non-decreasing sequences $(\boldsymbol{\alpha}, \boldsymbol{\beta})$
  - U: the length of the alignment
  - $\alpha_t \in \{1, \cdots, T_X\}$ and $\beta_t \in \{1, \cdots, T_Y\}$ for all $t \in \{1, \cdots, U\}$
- Given any local distance $d(x, y)$, the distance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as

$$D_{\boldsymbol{A}}^{(\boldsymbol{X},\boldsymbol{Y})} = \sum_{t=1}^{U} d(X_{\alpha_t}, Y_{\beta_t})$$

  - For instance, $d(x, y)$ can be squared Euclidean distance $\|x - y\|_2^2$



Expected Alignment

- Dynamic time warping (MDTW) takes one single best alignment from all possible alignments

$$D_{DTW}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{A} \in \mathcal{A}} D_{\boldsymbol{A}}^{(\boldsymbol{X},\boldsymbol{Y})}$$

  × Not satisfy triangle inequality

  × Training with local distance is non-differentiable

- The proposed distance on **expected alignment** takes the average distance over all possible alignment paths with a proper length $U \in [U_l, U_h]$

$$D_{EA}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}_{U \in [U_l, U_h]} \left[\mathbb{E}_{\boldsymbol{A} \in \mathcal{A}_U} D_{\boldsymbol{A}}^{(\boldsymbol{X},\boldsymbol{Y})} \right]$$

  ✓ Theoretical guarantees exist on metric validity

  ✓ Training and calculating can be simple and efficient

- A simple *sampling-based* method is designed to efficiently calculate the distance
  (a) Uniformly sample $U \in [U_l, U_h]$ as the alignment length
  (b) Uniformly sample an alignment of length U from all possible alignments

Learning Local Distance via Deep Networks

- We use multi-layer feed-forward network as the transformation function at each frame
  - Network weights are shared across different time steps
- Local distance is defined as the squared Euclidean distance of the transformed vectors
- In practice, 2-hidden-layer network with ReLU sigmoid activations works fine enough on our datasets

Learning DECADE via Large Margin Method

- Input: a set of time series $\{\boldsymbol{X}_i\}_{i=1}^N$ and their labels $\{Y_i\}_{i=1}^N$
- Overall objective function to minimize:

$$\mathcal{L}(D) = \sum_{i=1}^{N} \sum_{j \in \mathcal{S}_i^+} D^{(i,j)} + \lambda \sum_{i=1}^{N} \sum_{j \in \mathcal{S}_i^+} \sum_{k \in \mathcal{S}_i^-} \left[\delta + D^{(i,j)} - D^{(i,k)} \right]_+ + \mathcal{R}(D)$$

  - $\mathcal{L}^+(D)$: Reduce the distance of two time series with the same label
  - $\mathcal{L}^-(D)$: Increase the distance of two time series with different labels
  - $\mathcal{R}(D)$: Regularizations on our model. E.g., L2 loss on network weights, etc.

Theoretical Results on DECADE

- **Theorem 1.** *(Guarantees on the validity of DECADE)* When the local similarity measure $d(\boldsymbol{X}_t, \boldsymbol{Y}_{t'})$ is a valid distance metric, the expected alignment produces a valid pseudo-metric $D_{\text{EA}}(\boldsymbol{X}, \boldsymbol{Y})$. Namely, it satisfies all the three following properties:
  (a) $D_{\text{EA}}(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ (non-negativity)
  (b) $D_{\text{EA}}(\boldsymbol{X}, \boldsymbol{Y}) = D_{\text{EA}}(\boldsymbol{Y}, \boldsymbol{X})$ (symmetry)
  (c) $D_{\text{EA}}(\boldsymbol{X}, \boldsymbol{Y}) + D_{\text{EA}}(\boldsymbol{Y}, \boldsymbol{Z}) \geq D_{\text{EA}}(\boldsymbol{X}, \boldsymbol{Z})$ (triangle inequality)
- **Theorem 2.** *(Efficiency of the sampling method)* Given any two time series $\boldsymbol{X}$ and $\boldsymbol{Y}$ and the local distance is bounded by 1, if we approximate expected alignments with $\mathcal{O}\left(\frac{U_h^2}{\varepsilon^3}\right)$ alignment samples, with high probability we have

$$\left| D_{\text{EA}}(\boldsymbol{X}, \boldsymbol{Y}) - \hat{D}_{\text{EA}}(\boldsymbol{X}, \boldsymbol{Y}) \right| \leq \varepsilon$$

Quantitative Results

- Summary of 3 real-world datasets

Dataset	# of time series	# of time steps	# of features	# of classes	Prediction task
EEG	436	16	64	6	Alcoholic and # of stimuli
PhysioNet	918	48	17	2	In-hospital mortality
ICU	1734	24 - 36	13	2	In-hospital mortality

- DECADE achieves the best 1-nn classification accuracy on 2 of the 3 datasets

Method \ Dataset	EEG	PhysioNet	ICU
MDTW	0.3026 ± 0.06	0.6509 ± 0.05	0.7180 ± 0.02
GAK	0.3114 ± 0.05	0.6479 ± 0.05	0.6910 ± 0.03
MSA	0.2700 ± 0.03	0.6553 ± 0.05	0.6996 ± 0.02
ML-TSA	0.3375 ± 0.06	0.6406 ± 0.04	0.7123 ± 0.02
LDMLT-TS	0.3475 ± 0.03	0.6499 ± 0.04	$\mathbf{0.7278 \pm 0.03}$
MaLSTM	0.2963 ± 0.02	0.6886 ± 0.03	0.6926 ± 0.02
MSA-NN	0.3271 ± 0.05	0.6557 ± 0.02	0.7123 ± 0.02
MDTW-NN	0.3067 ± 0.05	0.6981 ± 0.02	0.7220 ± 0.02
DECADE	$\mathbf{0.3652 \pm 0.01}$	$\mathbf{0.7060 \pm 0.02}$	0.7232 ± 0.02

- Learning local distance and using expected alignment are two dispensable components for better performance

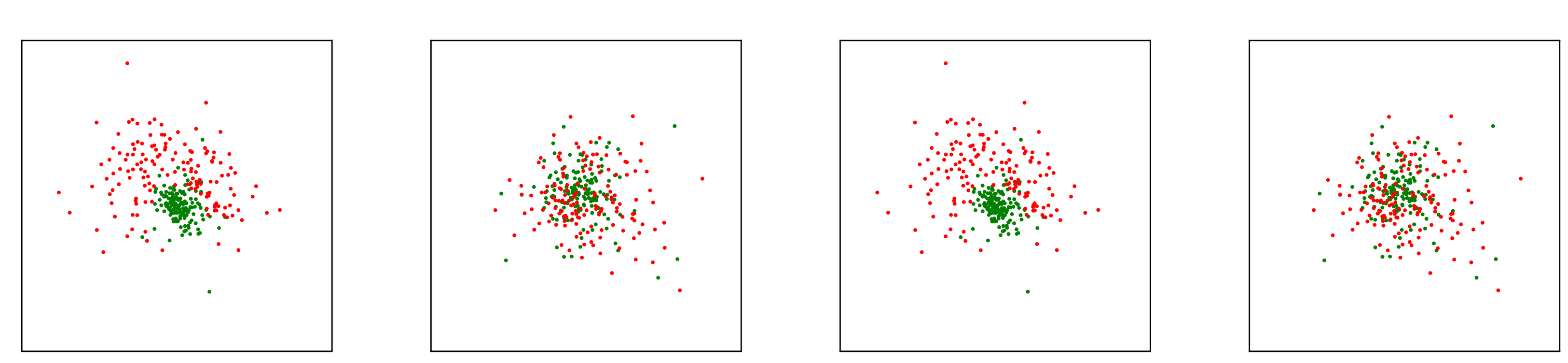
EEG		PhysioNet		ICU	
MDTW	*EA*[3]	*MDTW*	*EA*	*MDTW*	*EA*
0.3026 ± 0.06	0.2845 ± 0.03	0.6509 ± 0.05	0.5326 ± 0.05	0.7180 ± 0.02	0.6811 ± 0.01
MDTW-NN[4]	*DECADE*	*MDTW-NN*	*DECADE*	*MDTW-NN*	*DECADE*
0.3067 ± 0.05	$\mathbf{0.3652 \pm 0.01}$	0.6981 ± 0.02	$\mathbf{0.7060 \pm 0.02}$	0.7220 ± 0.02	$\mathbf{0.7232 \pm 0.02}$

[3] *EA: expected alignment + fixed L_2 local distance;* [4] *MDTW-NN: MDTW + learnable local distance.*

  - Learning data-dependent local distance always helps
  - MDTW performs better than EA without metric learning
  - DECADE achieves larger improvement than MDTW-NN by learning the data-dependent local distance

Visualization

- Embedding of PhysioNet dataset in 2 dimensions by multi-dimensional scaling (MDS) with learned pairwise distance *(Red: Patients with in-hospital mortality; Green: Live patients)*



DECADE **MDTW** **LDMLT-TS** **MaLSTM**

  - DECADE provided more coherent clusters of patients
  - Patients with in-hospital mortality (usually with extreme/abnormal values) spread out while live patients centered in the middle

References

[Berndt, James, 1994] Berndt, Donald J., and James Clifford. "Using dynamic time warping to find patterns in time series." KDD workshop 1994.

[Cuturi et al., 2007] Cuturi, Marco, et al. "A kernel for time series based on global alignments." ICASSP 2007.

[Hogeweg, Ben, 1984] Hogeweg, Paulien, and Ben Hesper. "The alignment of sets of sequences and the construction of phyletic trees: an integrated method." JME 1984.

[Garreau et al., 2014] Garreau, Damien, et al. "Metric learning for temporal sequence alignment." NIPS 2014.

[Mei et al., 2016] Mei, Jiangyuan, et al. "Learning a mahalanobis distance-based dynamic time warping measure for multivariate time series classification." CYB 2016.

[Mueller, Aditya, 2016] Mueller, Jonas, and Aditya Thyagarajan. "Siamese Recurrent Architectures for Learning Sentence Similarity." AAAI 2016.

[Hoeffding, 1963] Hoeffding, Wassily. "Probability inequalities for sums of bounded random variables." JASA 1963.