ICML | 2019
Thirty-sixth International Conference on Machine Learning

Learning Distance for Sequences by Learning a Ground Metric

Bing Su

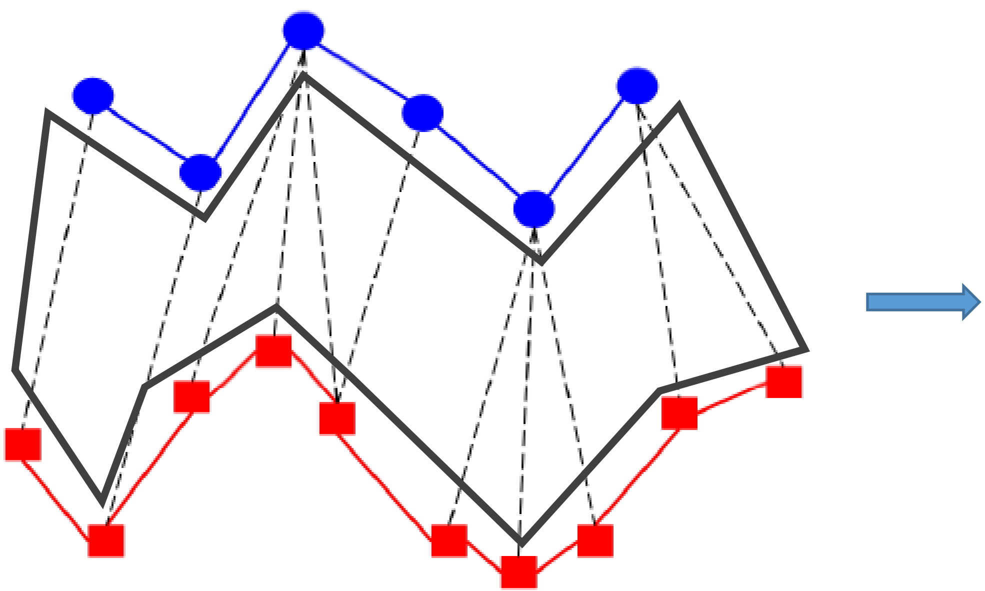
Ying Wu

Motivation

- Distance between sequences depends on ***temporal alignment*** to eliminate the local temporal discrepancies.

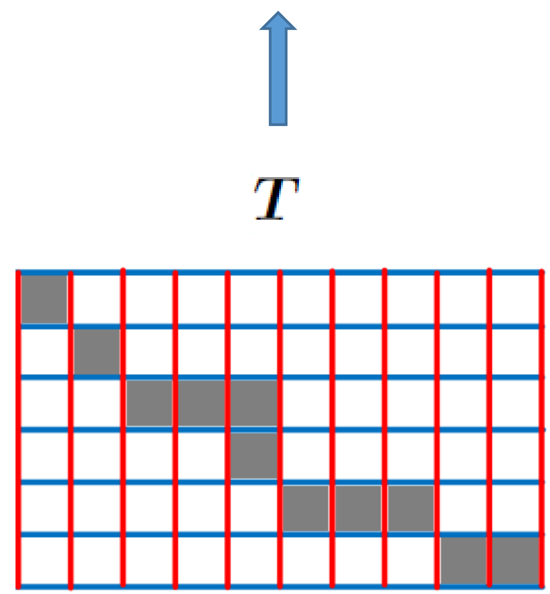
Temporal alignment

$$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{L_X}] \in \Omega^{L_X}$$



$$\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{L_Y}] \in \Omega^{L_Y}$$

$t_{i,j} = \boldsymbol{T}(i,j)$ indicates whether or the probability of the pair $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ is aligned.



Motivation

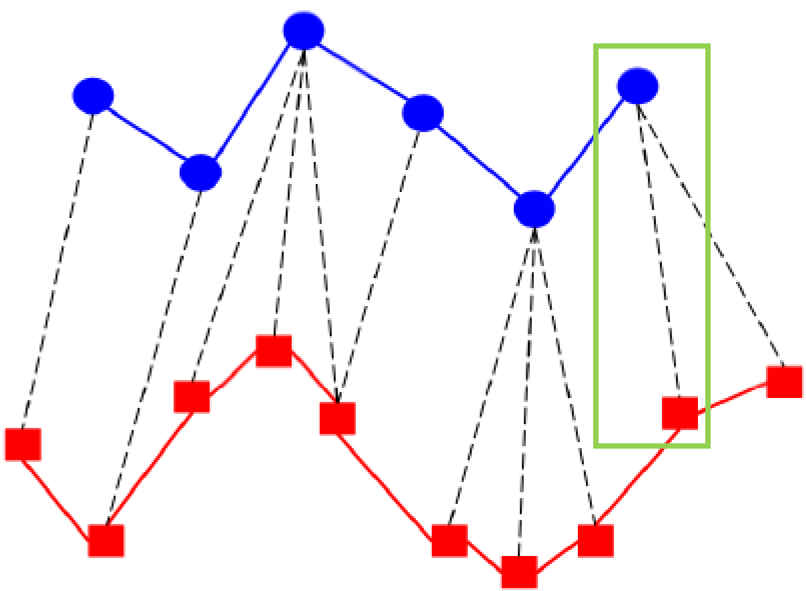
- The inference of alignment depends on the ***ground metric*** between elements in sequences.

Let Ω be a space, $d(M) : \Omega \times \Omega \to \mathbb{R}$ be the metric on this space.

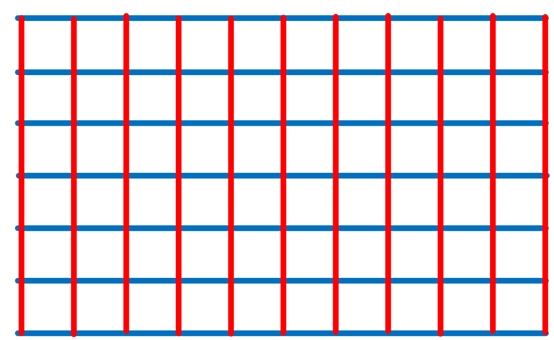
Ground metric

$$\boldsymbol{D}(M) := [d(\boldsymbol{M}, \boldsymbol{x}_i, \boldsymbol{y}_j)]_{ij} \in \mathbb{R}^{L_X \times L_Y}$$

$$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{L_X}] \in \Omega^{L_X}$$



$$\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{L_Y}] \in \Omega^{L_Y}$$



A Unified Perspective

- Distance between two sequences: a general formulation

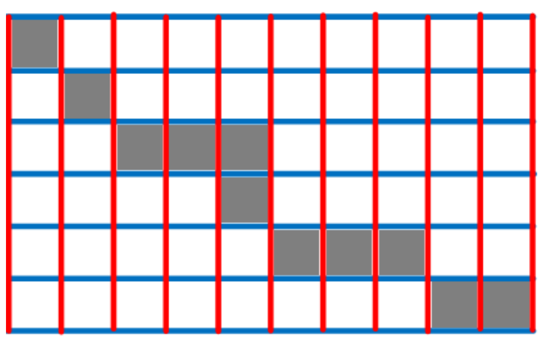
$$g_{\boldsymbol{M}}(\boldsymbol{X}, \boldsymbol{Y}) = \langle \boldsymbol{T}^*, \boldsymbol{D}(\boldsymbol{M}) \rangle$$

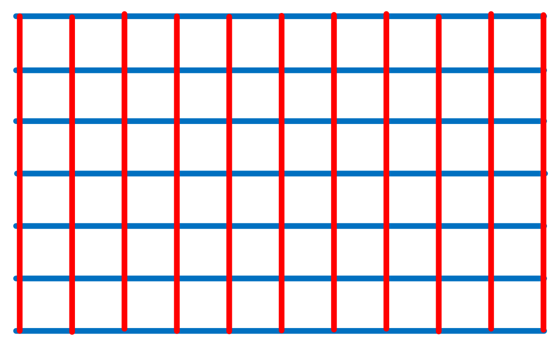
$\boldsymbol{T}$

The ***temporal alignment*** matrix

$\boldsymbol{D}(\boldsymbol{M})$

The ***ground metric*** matrix of pairwise distances between elements





A Unified Perspective

- ***T**** is generally inferred by

$$\boldsymbol{T}^* = \underset{\boldsymbol{T} \in \boldsymbol{\Phi}}{arg\min} \langle \boldsymbol{T}, \boldsymbol{D}(\boldsymbol{M}) \rangle + \mathscr{R}(\boldsymbol{T})$$

- $\boldsymbol{\Phi}$ is the feasible set of ***T***, which is a subset of $\mathbb{R}^{L_X \times L_Y}$ with some constraints; $\mathscr{R}(\boldsymbol{T})$ is a regularization term.
- Different distance measures for sequences differ in the constraints imposed to the feasible set, the regularization term, and the optimization method.

A Unified Perspective

- Connection to ***dynamic time warping (DTW)***

$$\mathscr{R}(\boldsymbol{T}) = 0;$$
$$\boldsymbol{\Phi} = \{\boldsymbol{T} \in \{0,1\}^{L_X \times L_Y} | \boldsymbol{T}_{1,1} = 1, \boldsymbol{T}_{L_X,L_Y} = 1;\ \boldsymbol{T}\boldsymbol{1}_{L_Y} > \boldsymbol{0}_{L_X}, \boldsymbol{T}^T\boldsymbol{1}_{L_X} > \boldsymbol{0}_{L_Y};$$
$$if\ t_{i,j} = 1, then\ t_{i-1,j+1} = 0, t_{i+1,j-1} = 0,\ \forall 1 < i < L_X, 1 < j < L_Y\}$$

DTW infers ***T*** via dynamic programming.

- Connection to ***order-preserving Wasserstein distance (OPW)***

$$\mathscr{R}(\boldsymbol{T}) = \lambda_1 I(\boldsymbol{T}) + \lambda_2 KL(\boldsymbol{T}||\boldsymbol{P});$$
$$\boldsymbol{\Phi} = \{\boldsymbol{T} \in \mathbb{R}_+^{L_X \times L_Y} | \boldsymbol{T}\boldsymbol{1}_{L_Y} = \tfrac{1}{L_X}\boldsymbol{1}_{L_X},\ \boldsymbol{T}^T\boldsymbol{1}_{L_X} = \tfrac{1}{L_Y}\boldsymbol{1}_{L_Y}\}$$

OPW infers ***T*** by the Sinkhorn's matrix scaling algorithm.

Problem

- The distance between sequences is formulated as a function of the ground metric: ***meta-distance***
- ***Learn meta-distance by learning the ground metric***
- Given a set of *N* training sequences and the corresponding labels, $\{\boldsymbol{X}^n, z^n\}_{n=1}^N$ $\boldsymbol{X}^n = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{L^n}] \in \mathbb{R}^{b \times L^n}$
- Learn a meta-distance $g_{\boldsymbol{M}}(\boldsymbol{X}^n, \boldsymbol{X}^{n'})$ by learning a Mahalanobis distance as the ground metric:

$$d(\boldsymbol{M}, \boldsymbol{x}_i, \boldsymbol{y}_j) = (\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{y}_j)$$

- $\boldsymbol{M} = \boldsymbol{W}\boldsymbol{W}^T$, $\boldsymbol{W} \in \mathbb{R}^{b \times b'}$
- Goal: with the learned ***W***, the resulting meta-distance

$$g_{\boldsymbol{M}}(\boldsymbol{X}^n, \boldsymbol{X}^{n'}) = g_{\boldsymbol{I}}(\boldsymbol{W}^T \boldsymbol{X}^n, \boldsymbol{W}^T \boldsymbol{X}^{n'})$$

better separates sequences from different classes.

Objective

- ***Regressive virtual sequence metric learning (RVSML)***
- Associate a virtual sequence $\boldsymbol{V}^n = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{l^n}] \in \mathbb{R}^{b' \times l^n}$ with each training sequence $\boldsymbol{X}^n$
- Minimize the meta-distances between the training sequences and their associated virtual sequences

$$\min_{\boldsymbol{W}} \frac{1}{N} \sum_{n=1}^{N} g_{\boldsymbol{I}}(\boldsymbol{W}^T \boldsymbol{X}^n, \boldsymbol{V}^n) + \beta \|\boldsymbol{W}\|_{\mathcal{F}}^2$$
$$= \frac{1}{N} \sum_{n=1}^{N} \langle \boldsymbol{T}^{n*}, \boldsymbol{D}_{\boldsymbol{I}}^n(\boldsymbol{W}) \rangle + \beta \|\boldsymbol{W}\|_{\mathcal{F}}^2$$
$$s.t.\ \boldsymbol{T}^{n*} = arg \min_{\boldsymbol{T} \in \boldsymbol{\Phi}} \langle \boldsymbol{T}^n, \boldsymbol{D}_{\boldsymbol{I}}^n(\boldsymbol{W}) \rangle + \mathscr{R}(\boldsymbol{T}^n)$$

- If $\mathscr{R}(\boldsymbol{T})$ does not depend on ***W***, it is equivalent to

$$\min_{\boldsymbol{W}, \boldsymbol{T}^n} \frac{1}{N} \sum_{n=1}^{N} \langle \boldsymbol{T}^n, \boldsymbol{D}_{\boldsymbol{I}}^n(\boldsymbol{W}) \rangle + \beta \|\boldsymbol{W}\|_{\mathcal{F}}^2 + \mathscr{R}(\boldsymbol{T}^n)$$

Optimization

$$\min_{\boldsymbol{W},\boldsymbol{T}^n} \frac{1}{N}\sum_{n=1}^{N}\langle \boldsymbol{T}^n, \boldsymbol{D}_{\boldsymbol{I}}^n(\boldsymbol{W})\rangle + \beta\|\boldsymbol{W}\|_{\mathcal{F}}^2 + \mathscr{R}(\boldsymbol{T}^n)$$

- Fix $\boldsymbol{T}^n$, optimize ***W***: standard regression, closed form solution

$$\boldsymbol{W}^* = \boldsymbol{A}^{-1}\left(\sum_{n=1}^{N}\sum_{i=1}^{L^n}\sum_{j=1}^{l^n} t_{ij}^n \boldsymbol{x}_i^n {\boldsymbol{v}_j^n}^T\right) \qquad \boldsymbol{A} = \sum_{n=1}^{N}\sum_{i=1}^{L^n}\sum_{j=1}^{l^n} t_{ij}^n \boldsymbol{x}_i^n {\boldsymbol{x}_i^n}^T + \beta N \boldsymbol{I}$$

- Fix ***W***, optimize $\boldsymbol{T}^n$: standard inference, e.g. DTW, OPW

$$\boldsymbol{T}^{n*} = arg \min_{\boldsymbol{T}^n \in \boldsymbol{\Phi}} \langle \boldsymbol{T}^n, \boldsymbol{D}_{\boldsymbol{I}}^n(\boldsymbol{W})\rangle + \mathscr{R}(\boldsymbol{T}^n)$$

- Guaranteed convergence

Evaluation

- Generating $\boldsymbol{V}^n = f(\boldsymbol{X}^n, z^n) = [\boldsymbol{e}_{(z^n-1)m+1}, \cdots, \boldsymbol{e}_{(z^n-1)m+m}]$
- RVSML instantiated by (a) DTW and (b) OPW using the NN classifier with the (a) DTW and (b) OPW distance
- Comparison with other metric learning methods on the ChaLearn and SAD datasets

(a) DTW

Method	MAP	Accuracy
Ori (Su & Hua, 2018)	11.75	61.12
ITML (Davis et al., 2007)	13.46	52.17
LMNN (Weinberger & Saul, 2009)	11.67	63.78
RVML (Perrot & Habrard, 2015)	31.21	83.79
LDMLT (Mei et al., 2014)	21.30	84.37
SWMD (Huang et al., 2016)	14.39	64.45
RVSML	**33.83**	**87.38**

(b) OPW

Method	MAP	Accuracy
Ori (Su & Hua, 2018)	12.21	59.38
ITML (Davis et al., 2007)	13.92	64.71
LMNN (Weinberger & Saul, 2009)	12.07	62.83
RVML (Perrot & Habrard, 2015)	30.19	80.66
LDMLT (Mei et al., 2014)	21.56	82.74
SWMD (Huang et al., 2016)	15.36	60.31
RVSML	**33.07**	**83.82**

(a) DTW

Method	MAP	Accuracy
Ori (Su & Hua, 2018)	56.58	96.36
ITML (Davis et al., 2007)	51.13	95.55
LMNN (Weinberger & Saul, 2009)	56.25	96.00
SCML (Shi et al., 2014)	47.98	93.27
RVML (Perrot & Habrard, 2015)	57.94	**96.59**
LDMLT (Mei et al., 2014)	59.54	96.50
SWMD (Huang et al., 2016)	52.44	93.95
RVSML	**60.24**	96.23

(b) OPW

Method	MAP	Accuracy
Ori (Su & Hua, 2018)	59.77	96.36
ITML (Davis et al., 2007)	54.51	96.36
LMNN (Weinberger & Saul, 2009)	59.33	96.27
SCML (Shi et al., 2014)	50.08	94.50
RVML (Perrot & Habrard, 2015)	60.71	95.77
LDMLT (Mei et al., 2014)	61.07	96.73
SWMD (Huang et al., 2016)	58.00	95.41
RVSML	**65.63**	**97.09**

Results

- Comparison with state-of-the-art methods on the MSR Activity3D and MSR Action3D datasets

Method	Accuracy
Actionlet Ensemble (Wang et al., 2012)	85.8%
Moving Pose (Zanfir et al., 2013)	73.8%
COV-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014)	75.5%
Ker-RP-POL (Wang et al., 2015)	96.9%
Ker-RP-RBF (Wang et al., 2015)	96.3%
Kernelized-COV (Cavazza et al., 2016)	96.3%
Luo et al. (Luo et al., 2017)	86.9%
Ji et al. (Ji et al., 2018)	81.3%
DSSCA SSLM (Shahroudy et al., 2018)	**97.5%**
RVSML-DTW+Kernelized-COV	96.9%
RVSML-OPW+Kernelized-COV	**97.5%**

Method	Accuracy
Actionlet Ensemble (Wang et al., 2012)	88.2%
Moving Pose (Zanfir et al., 2013)	91.7%
COV-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014)	80.4%
Ker-RP-POL (Wang et al., 2015)	96.2%
Ker-RP-RBF (Wang et al., 2015)	**96.9%**
Kernelized-COV (Cavazza et al., 2016)	96.2%
SCK+DCK (Koniusz et al., 2016)	91.45%
TS-LSTM-GM (Lee et al., 2017)	91.21%
FTP-SVM (Ben Tanfous et al., 2018)	90.01%
Bi-LSTM (Ben Tanfous et al., 2018)	86.18%
RVSML-DTW+Kernelized-COV	82.78%
RVSML-OPW+Kernelized-COV	**96.34%**
RVSML-DTW+TS-LSTM-GM	93.04%
RVSML-OPW+TS-LSTM-GM	90.48%

- Please visit our poster for more details.
- Thank you very much!