Weakly supervised machine learning for time series

Romain Tavenard (Université de Rennes)
@rtavenar

TSdays
Rennes, March 26th, 2019
Weakly supervised machine learning for time series
Part I

**tslearn**: A Python ML toolkit for time series

Part II

Weakly supervised learning: why & how?

Part III

Beyond time series: structured data in general
Part I
tslearn: A Python ML toolkit for time series
tslearn: A Python ML toolkit for time series

• Install it (via pip or conda)
  pip install tslearn

• Play with it

>>> from tslearn.datasets import UCR_UEA_datasets
>>> from tslearn.clustering import TimeSeriesKMeans

>>> X_train, y_train, X_test, y_test = UCR_UEA_datasets().load_dataset("TwoPatterns")

>>> print(X_train.shape)
(1000, 128, 1)

>>> km = TimeSeriesKMeans(n_clusters=3, metric="dtw")
>>> km.fit(X_train)
tslearn: A Python ML toolkit for time series
Samples from the gallery of examples [1/2]

Link to online notebook
tslearn: A Python ML toolkit for time series
Samples from the gallery of examples [2/2]

Link to online notebook
tslearn: A Python ML toolkit for time series
Feel free to contribute

• All contributions are welcome (via github)
  • New feature requests
  • Bug reports
  • Bug fixes
  • Improved documentation

• Most contributors are volunteers
• 1-year dev funding scheduled in ANR MATS (applications welcome anytime)
Part II
Weakly supervised learning: why & how?
Remote sensing & ML

- **Labelled** training samples are...
  - Costly to acquire
  - Noisy
- Weakly supervised learning
  - Domain adaptation (not covered in this talk)
    - Across years
    - Across regions
    - Great stuff by Courty *et al.*
- Representation learning
  - **Self-supervised approaches**
Learning DTW-Preserving Shapelets (LDPS) [Lods et al., 2017]

- **Self-supervised learning:** generate supervised information from the data itself
- **Here:** Learn a Shapelet Transform to mimic DTW
Learning DTW-Preserving Shapelets (LDPS)
Dynamic Time Warping: a tool for time series analysis

- Dynamic Time Warping (DTW)
  - *Elastic* similarity measure
  - Invariant to time shifts
  - Costly to compute
Learning DTW-Preserving Shapelets (LDPS) 

\( k \)-means and Dynamic Time Warping

- **Issue:** \( k \)-means needs barycenters
  - Explicit formulation in Euclidean Space
  - No easy way for DTW

- **Option 1:** DBA \( k \)-means [Petitjean *et al.*, 2011]
- **Option 2:** Soft-DTW [Cuturi & Blondel, 2017]
Learning DTW-Preserving Shapelets (LDPS)
State-of-the-art: Time Series Shapelets

- Shapelets
  - **Discriminant** subseries
  - Learnt in [Grabocka et al., 2014]

- Shapelet Transform
  [Hills et al., 2014]
  - Efficient computation
  - Invariant to time shifts
Learning DTW-Preserving Shapelets (LDPS)
Schematic view of our model

1. First, compute the distance between $\tau^j_i$ and $s^k$ for each time series $i$ and each shapelet ($i=1, 2, ..., N$ and $k=1, 2, ..., K$) with $j \in \{1, 2, ..., L+1\}$ such that we compute the distance between all the possible subseries with each shapelet. Shapelets from LTS are equivalent to filters of a CNN convolution step, computing the distance between shapelets and each subsequence of a time series thus corresponds exactly to a convolution layer of a CNN. For LTS, the convolution exploits the temporal relationship between neighboring points, which is equivalent to a convolution applied on an image which exploits the spatial relationship between neighboring pixels. Indeed, the different subsequence correspond to smaller series extracted at regular step i.e. using a sliding window technique.

2. Then, we select the minimal distances between $x^i$ and $s^k$ for $i \in \{1, 2, ..., N\}$ and $k \in \{1, 2, ..., K\}$. This step corresponds to a min-pooling layer.

3. Finally, we use the weights of the logistic regression $w$ in order to predict the class label of the time series.

All these steps are pictured in Figure 3.6, where:

- The red and blue bullets respectively correspond to the distance between the first shapelet $s_1$ and $(\tau^1_i, \tau^2_i)$.
- The grey bullet corresponds to the minimal distance between the first shapelet $s_1$ and the $i$-th time series $x_i$.
- In order to get the predictions, $w^M + w^0$ is computed as in the LTS algorithm.

In the literature, there is – to our knowledge – only Cui et al. [2016]; Lods et al. [2017] that also mentions that LTS can be viewed as a special case of a CNN.
Learning DTW-Preserving Shapelets (LDPS)

Schematic view of our model

Learning DTW-Preserving Shapelets (LDPS)
Learning DTW-Preserving Shapelets (LDPS)
Problem formulation

- Loss function
\[
\mathcal{L}(\{T_i\}) \propto \sum_{i_1} \sum_{i_2} (\text{DTW}(T_{i_1}, T_{i_2}) - \beta \| m_\theta(T_{i_1}) - m_\theta(T_{i_2}) \|_2^2)^2
\]

- Optimize jointly on the mapping \( m_\theta \) and the scaling factor \( \beta \)
- Self-supervised learning
  - No need for human annotation
Learning DTW-Preserving Shapelets (LDPS)
Experiment #1: Quality of fit
Learning DTW-Preserving Shapelets (LDPS)

Experiment #2: Clustering using $k$-means
Learning DTW-Preserving Shapelets (LDPS)
Experiment #3: Retrieval [Sperandio et al., 2018]

- kNN search in a database
- DTW as a target metric (but too costly)
Learning DTW-Preserving Shapelets (LDPS)
Take-Away slide

- Learning DTW-Preserving Shapelets (Arnaud Lods)
  - embeds time series in a metric space
    - useful for many data analysis tasks
  - uses only self-supervision
- Perspectives
  - can be extended to any similarity measure (pick the best for your data!)
  - LDPS + semi-supervised learning: Learn a mix between DTW and a good classifier (mixed loss function)
Part III
Beyond time series: structured data in general
Beyond Time Series: Structured data in general
Example [1/2]: GPT-2 Language Model

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state of the art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization —

Twitter status, @openai, Feb 14th, 2019
Beyond Time Series: Structured data in general

Example [2/2]: Early classification of time series

M. Rußwurm et al.
End-to-end Learning for Early Classification of Time Series
Beyond Time Series: Structured data in general

DTW & Optimal Transport (OT)

DTW problem = OT problem + strong structural constraints

Dynamic Time Warping

Wasserstein distance
Beyond Time Series: Structured data in general
Optimal Transport & Structure

\[ d_{GW}(X, Y) = \min_{\pi} \sum_{(x, x') \in X \times X'} |d_X(x, x') - d_Y(y, y')| \]

Gromov-Wasserstein distance
Beyond Time Series: Structured data in general
Optimal Transport & Structure — cont’d

Wasserstein distance
$$\min_\pi \sum_{i,j} d_{ij} \pi_{i,j}$$

Gromov-Wasserstein distance
$$\min_\pi \sum_{i,j,k,l} L_{ijkl} \pi_{i,j} \pi_{k,l}$$

Fused-Gromov-Wasserstein (FGW) distance
$$\min_\pi \sum_{i,j,k,l} \left[ (1 - \alpha) d_{ij} + \alpha L_{ijkl} \right] \pi_{i,j} \pi_{k,l}$$
Beyond Time Series: Structured data in general
FGW & time series barycenters

Euclidean barycenter \((N = 275)\)

Soft-DTW barycenter \((\gamma = 1, N = 20)\)

DBA barycenter \((N = 20)\)

FGW barycenter \((\alpha = 10^{-6}, N = 20)\)
Beyond Time Series: Structured data in general

Take-Away slide

- Fused Gromov-Wasserstein (Titouan Vayer)
  - is a proper distance metric
  - is differentiable
  - can be used for
    - clustering of structured data with explicit barycenters
    - structured data classification
    - …

- Structured data
  - is awesome for weakly supervised approaches (cf. GPT-2)
Part IV
If I had more time...
Learning Interpretable Shapelets through Adversarial Regularization
[Yichang Wang — Work In Progress]
Learning Interpretable Shapelets through Adversarial Regularization [Yichang Wang — Work In Progress]

Figure 4: Illustration of the evolution of a shapelet during training (for the Wine dataset).
Interpretable Shapelets
[Yichang Wang — Work In Progress]

(a) Learning Shapelets [9]          (b) AI↔PR-CNN
Conclusion & Perspectives

• `tslearn` needs you ;-)  
  • Or, at least, OSS needs you
• Representation learning
  • Learn to mimic a target metric
  • *Learn a representation from structural information (generative models)*
• Interpretable models for Time Series
  • Useful for unsupervised settings
References

• R. Tavenard. 
  *tslearn: A machine learning toolkit dedicated to time-series data.*
  https://github.com/rtavenar/tslearn

