An Interpretable Distance Measure for Multivariate Non-Stationary Physiological Signals

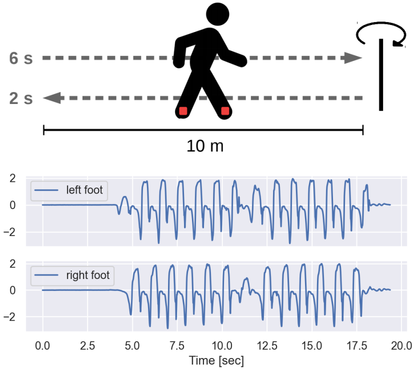
Sylvain W. Combettes, Charles Truong, and Laurent Oudre

Centre Borelli, ENS Paris-Saclay

ICDM AI4TS workshop – Shanghai, China – December 1st, 2023

I) Context and motivation

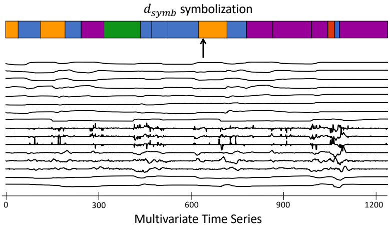
I.1) Comparing multivariate non-stationary physiological signals



Motivation: study of human locomotion [3]

- Angular velocity recorded on the left and right feet using a pair of sensors.
- Protocol: standing, walking, turning around, walking back, and standing.
- Multivariate signals with $d = 16$ dimensions: norms of the STFT (Short Time Fourier Transform) of each foot recording (univariate signal).

I.2) Our approach: symbolization, then distance on strings



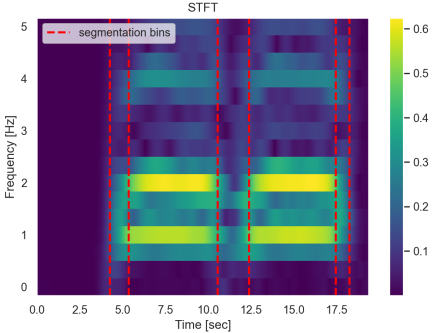
- Popular distances between multivariate time series (Euclidean distance, Dynamic Time Warping) can not handle non-stationarity.
- Our distance is interpretable and can compare non-stationary signals: (i) symbolization, (ii) distance on strings.

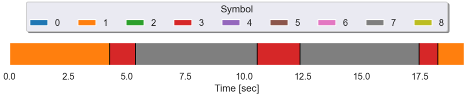
Symbolization technique

1. Segmentation step: a real-valued signal of length n is split into w segments ($w < n$).
2. Quantization step: each segment is mapped to a discrete value taken from a set of A symbols.
   Example of set of symbols with $A = 5$: $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}\}$.

II) The d_symb symbolization and distance measure

II) The d_symb symbolization and distance measure





Steps of d_{symb}

1. Segmentation: change-point detection (on the mean).
2. Quantization: K-means clustering (of the means per segment), with $K = A$.
3. Distance: general edit distance between the resulting symbolic signals.

II.1) Segmentation

Change-point detection: finding the w^* unknown instants $t_1^* < t_2^* < \ldots < t_{w^*+1}^*$ where the mean of signal $x = (x_1, \ldots, x_n)$ change abruptly:

$$\left(\hat{w}, \hat{t}_1, \ldots, \hat{t}_{\hat{w}+1}\right) = \underset{(w, t_1, \ldots, t_{w+1})}{\arg\min} \sum_{k=0}^{w+1} \sum_{t=t_k}^{t_{k+1}-1} \left\| x_t - \bar{x}_{t_k:t_{k+1}} \right\|^2 + \lambda w, \qquad (1)$$

where $\bar{x}_{t_k:t_{k+1}}$ is the empirical mean of $\{x_{t_k}, \ldots, x_{t_{k+1}-1}\}$ and $\lambda > 0$ is a penalization parameter.

Remarks

- Compromise between the reconstruction error and the number of change-points.
- When λ is small, many change-points are detected.
  For calibration purposes, we use $\lambda = \ln(n)$ [4].
- Solved using the Pruned Exact Linear Time (PELT) algorithm [1], which is shown to have $\mathcal{O}(n)$ complexity (under some assumptions).

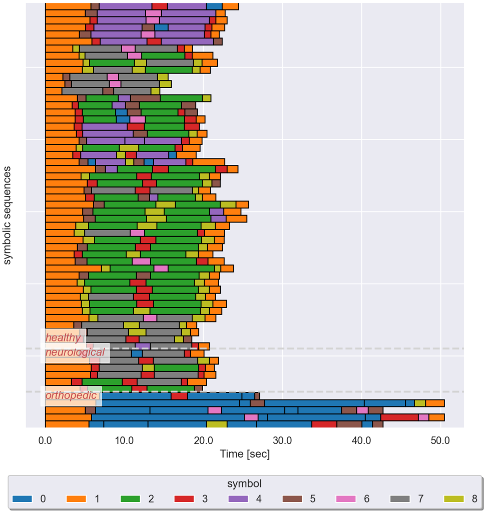
II.2) Distance measure

The d_{symb} distance measure: levering the general edit distance

1. Preprocessing.
   - Including the segment length information: replicating each symbol proportionally to its segment length.
     Example: `abd` becomes `aabbbbdd`.
   - Shortening: dividing each length by the minimum length.
     Example: `aabbbbdd` becomes `abbd`.
2. Applying the general edit distance with custom costs.
   - Edit distance on strings (a.k.a Levenshtein distance [2]): minimal cost of a sequence of operations that transform a string into another.
   - Allowed simple operations and their costs:
     - Substitution: Euclidean distance between the cluster centers of the symbols.
     - Insertion: max of substitution costs.
     - Deletion: max of substitution costs.
   - Total cost: sum of the costs of the simple operations.

III) Experimental results

III.1) Interpretation of the d_symb symbolization



Color bars for 60 recordings.

Observations

- The general structure is coherent with the protocol.
- Change-point detection finds stationary segments.
- Each symbol can be associated with a type of behavior.

III.2) Interpretation of the d_symb distance measure

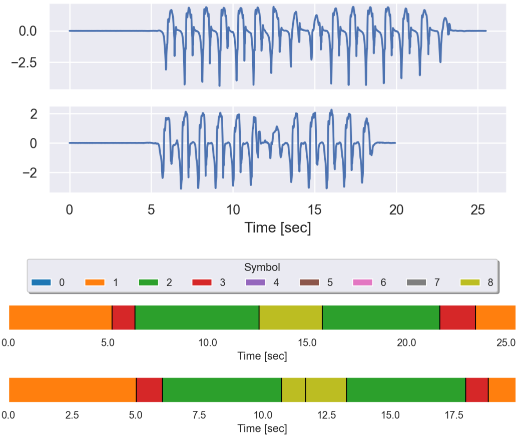
Benchmark: computing the silhouette score

- We have 3 groups of patients:
  - healthy,
  - neurological pathology (such as cerebellar disorder),
  - or orthopedic pathology (such as knee injuries).
- The silhouette coefficient is calculated using the distance matrix and the ground truth labels corresponding to the patient group.

Distance measure	Mean Silhouette score	Median Silhouette score
DTW-D	0.15	0.18
DTW-I	0.15	0.19
d_{symb}	0.33	0.40

III.2) Interpretation of the d_symb distance measure

Robustness to the difference in length



Observations

- The two scaled univariate gait signals are different in length...
- but are considered similar by d_{symb} (applied to their multivariate spectrograms).

Thank you for your attention.

sylvain.combettes@ens-paris-saclay.fr

https://sylvaincom.github.io

References

R. Killick, P. Fearnhead, and I. A. Eckley.
Optimal detection of changepoints with a linear computational cost.
Journal of the American Statistical Association, 107(500):1590–1598, 2012.

V. I. Levenshtein et al.
Binary codes capable of correcting deletions, insertions, and reversals.
In Soviet Physics Doklady, volume 10, pages 707–710, 1966.

C. Truong, R. Barrois-Müller, T. Moreau, C. Provost, A. Vienne-Jumeau, A. Moreau, P.-P. Vidal, N. Vayatis, S. Buffat, A. Yelnik, D. Ricard, and L. Oudre.
A Data Set for the Study of Human Locomotion with Inertial Measurements Units.
Image Processing On Line, 9:381–390, 2019.
https://doi.org/10.5201/ipol.2019.265.

References

C. Truong, L. Oudre, and N. Vayatis.
Selective review of offline change point detection methods.
Signal Processing, 167:107299, 2020.