d_{symb} playground: an interactive tool to explore large multivariate time series datasets

Sylvain W. Combettes, Paul Boniol, Charles Truong, Laurent Oudre

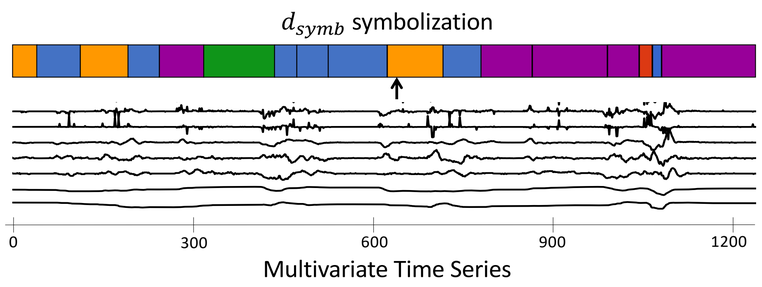
Université Paris-Saclay, Université Paris Cité, ENS Paris-Saclay, CNRS, SSA, INSERM, Centre Borelli, Inria, ENS

Motivation

- Sensors are increasingly used in various fields. These signals can be multivariate and non-stationary [1].
- Symbolization transforms a real-valued signal of length n into a discrete-valued signal of smaller length $w < n$, called a symbolic sequence.
- Popular distances such as the Euclidean distance and Dynamic Time Warping can not handle non-stationary time series [2].

d_{symb} symbolization and distance measure

Symbolization and distance measure for multivariate time series [2]. d_{symb} requires one parameter (the number of symbols A) and is composed of three steps:



(1) Segmentation

Change-point detection (on the mean).

Parameter λ ($\lambda = \ln(n)$ by default) to control the number of change-points.

Solved with PELT algorithm [3] with $O(n)$ complexity.

(2) Quantization

K-means clustering (of the means per segment).

We set the number of clusters K to the desired number of symbols A.

Replication of each symbol proportionally to its segment length.

(3) Distance

General edit distance between the symbolic representations with custom costs:

Substitution: *Euclidean distance between the centroids of the symbols*

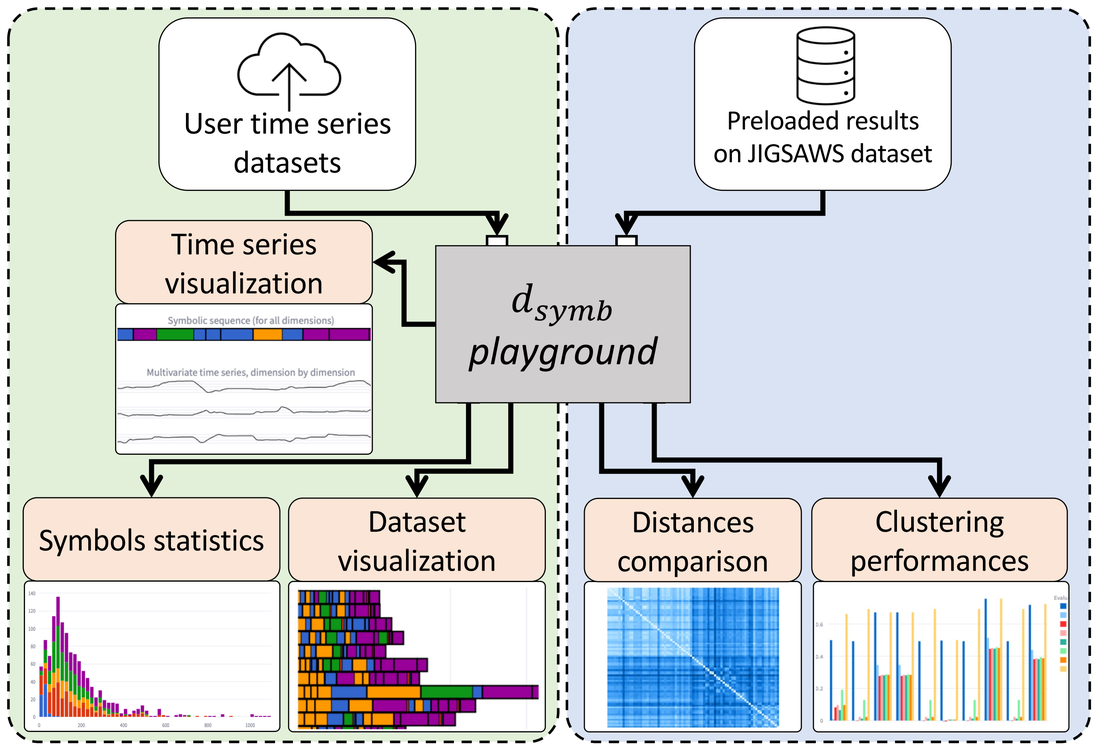
Insertion/Deletion: *max of substitution costs*

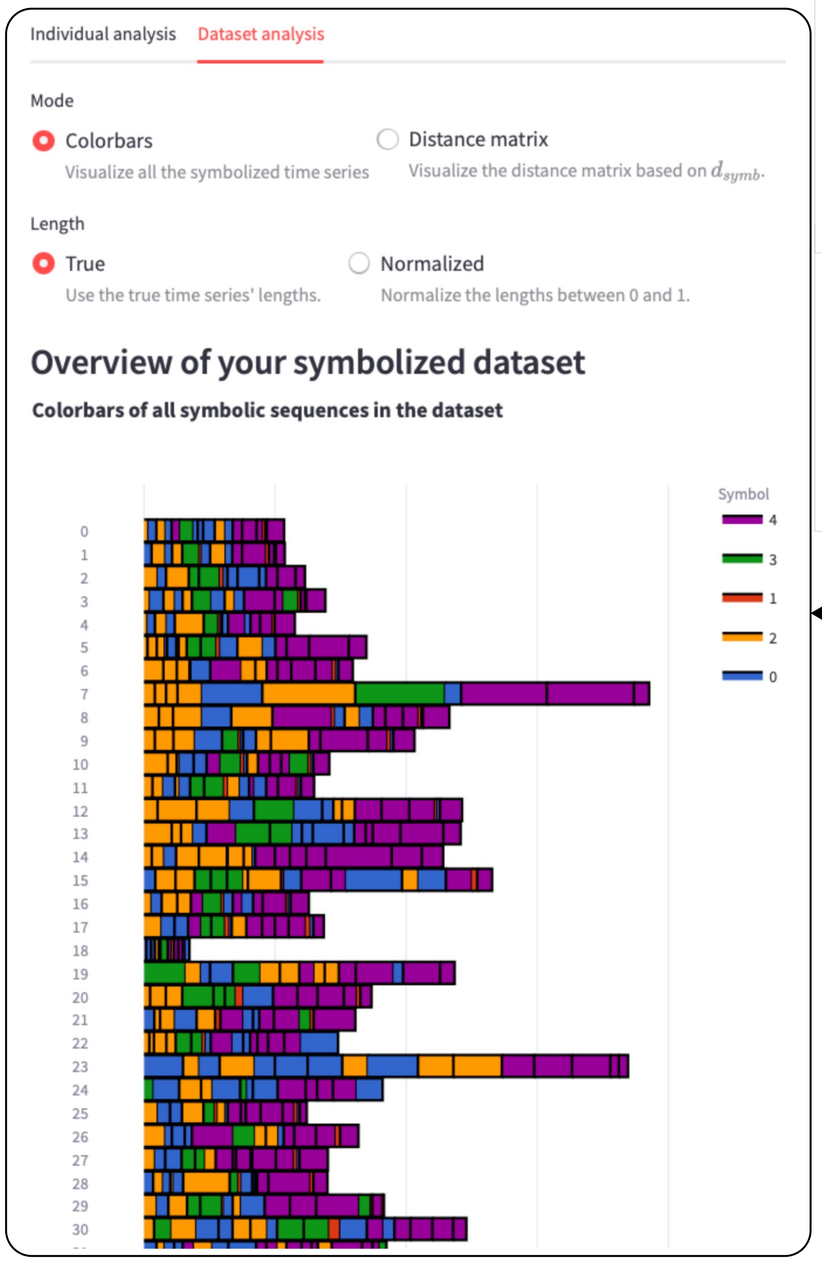
d_{symb} Playground

Stand-alone webapp developed using Streamlit proposing three main features:

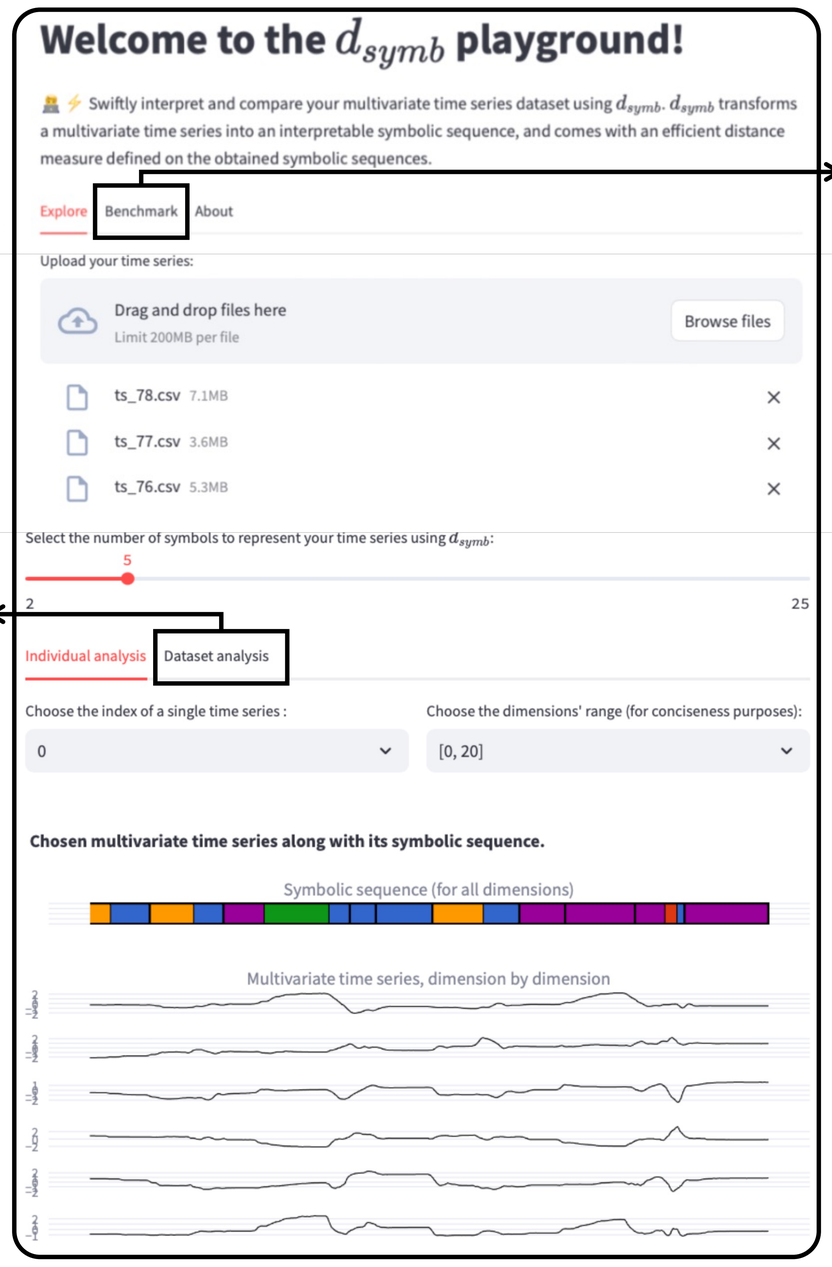
1. Visualizing uploaded time series and their symbolizations
2. Comparing time series based on their symbolizations
3. Benchmarking distance measures

Input/Features summary

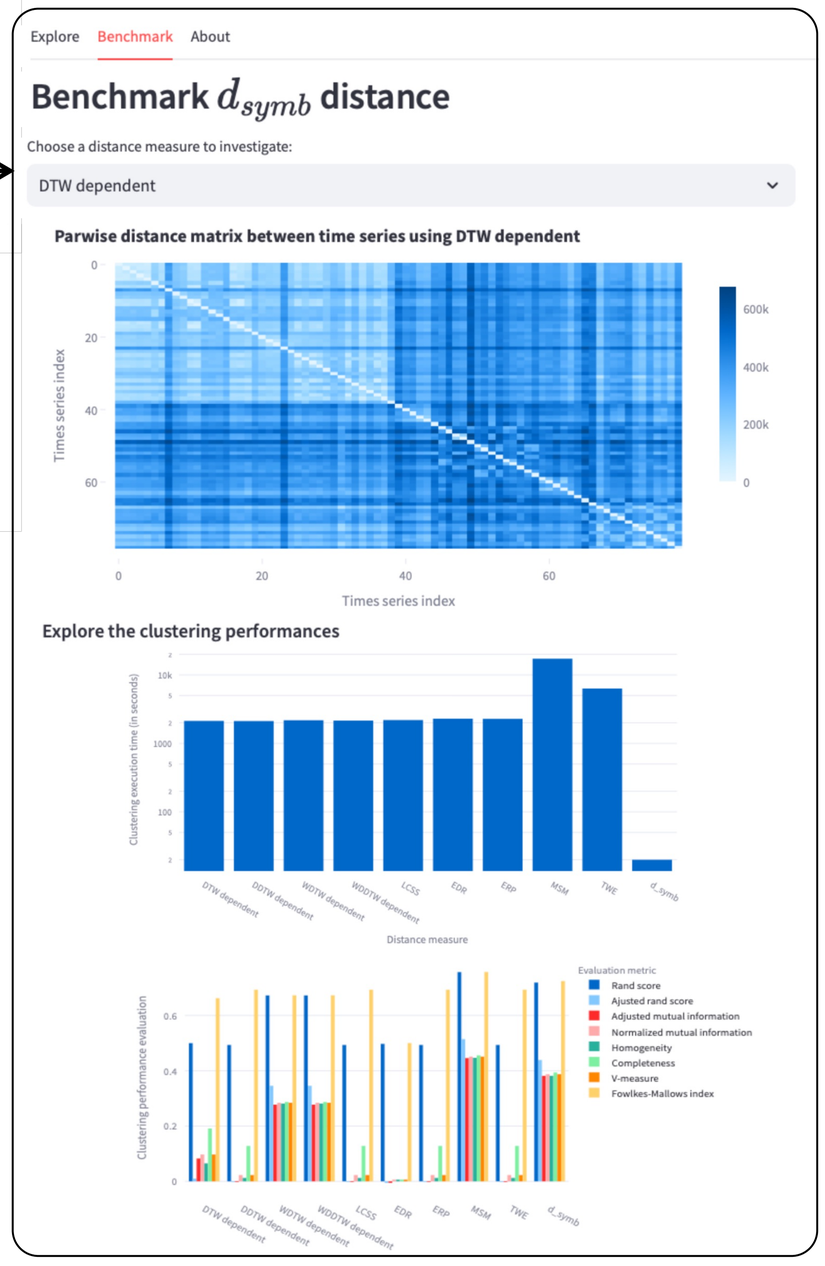




(A) Dataset analysis frame



(B) Individual analysis frame



(C) Benchmark frame

Bibliography

[1] Sylvain W. Combettes, Paul Boniol, Antoine Mazarguil, Danping Wang, Diego Vaquero-Ramos, Marion Chauveau, Laurent Oudre, Nicolas Vayatis, Pierre-Paul Vidal, Alexandra Roren, and Marie-Martine Lefèvre-Colau, Arm-CODA: A Data Set of Upper-limb Human Movement During Routine Examination, Image Processing On Line, 14 (2024), pp. 1–13.

[2] S. W. Combettes, C. Truong, and L. Oudre. An interpretable distance measure for multivariate non-stationary physiological signals. In Proceedings of the International Conference on Data Mining (AI4TS Workshop), 2023.

[3] 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

d_{symb} playground Webapp

Github repository