CDPS: Constrained DTW-Preserving Shapelets

Hussein El Amouri* Thomas Lampert* Clement Mallet† Pierre Gançarski*

*ICube, University of Strasbourg, France. † LASTIG, Univ Gustave Eiffel, IGN, ENSG, France

Introduction and Motivation

Constrained Clustering

- Semi-supervised approach.
- It balances between the labeling effort and the expert expectation.
- The expert provides constraints to guide the algorithm to a result that is within his/her needs.

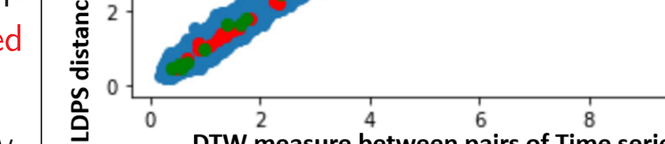
Time series (TS) Data

- Ordered set of observations, suffering from distortions in time and amplitude. Matching TS is often done using DTW similarity[a] instead of Euclidean distance.

[a]Senin, P., 2008. Dynamic time warping algorithm review.

Problem

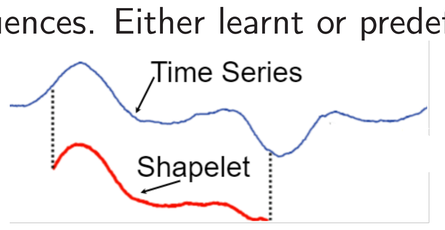
- Experts usually propose obvious constraints: no respect for Informativeness and/or Coherence.
  Measuring these properties requires a Euclidean space, which is not possible with DTW similarity.



- No approach for automatic, useful constraint proposal.

Shapelets

- Discriminative and interpretable time series sub-sequences. Either learnt or predefined.



- Shapelet Transform[b]: Maps time series T to vector from $\overline{T}$ in Euclidean space.

[b]Lines et. al: A shapelet transform for time series classification. In: SIGKDD. pp. 289–297 (2012)

Our contribution: CDPS- Constrained DTW Preserving Shapelets

We address the problem of measuring constraint properties by providing a Euclidean space that encapsulates the user constraints. The following shapelet loss is proposed:

$$\mathcal{L}(T_i, T_j) = \frac{1}{2}\left(DTW(T_i, T_j) - \beta||\overline{T}_i - \overline{T}_j||_2\right)^2 + \begin{cases} \alpha||\overline{T}_i - \overline{T}_j||_2^2, & \text{if } (i,j) \in \mathsf{ML}, \\ \gamma \max(0, w - ||\overline{T}_i - \overline{T}_j||_2^2, & \text{if } (i,j) \in \mathsf{CL}, \\ 0, & \text{otherwise.} \end{cases}$$

- First term inspired from (LDPS) Learning DTW Preserving Shapelets[c], ensures that DTW similarity between time series is preserved in the transformed space, see Figure 1.(a). In this setting no constraints are provided, we can observe that similar and dissimilar points are mixed up.
- Second term: Contrastive loss. Ensures cannot link samples are far in the transformed space and must link samples are close. This is reflected in figure 1.(b) where 25% of the samples are under constraints. Green points are more similar Red points are more dissimilar[d].

[c]Lods, A., Malinowski, S., Tavenard, R., Amsaleg, L.: Learning DTW-preserving shapelets. In: IDA (2017)

[d]CBF dataset is used, from the UCR archive.

(a) Learning with no Constraints (LDPS) (b) Learning with constraints (CDPS)

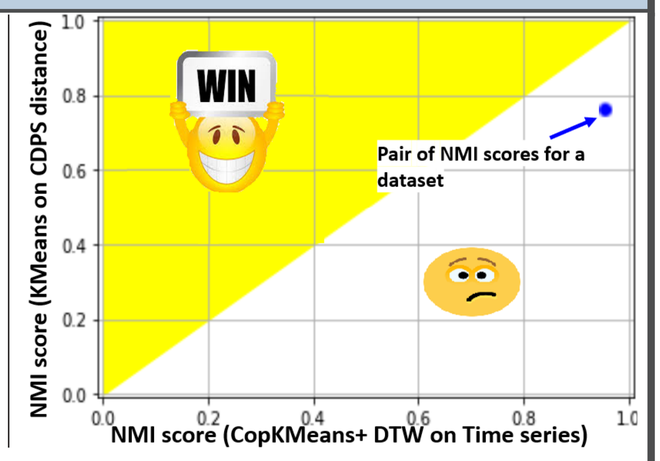
Figure 1: CDPS ensures CL samples are far apart from ML samples while LDPS fails.

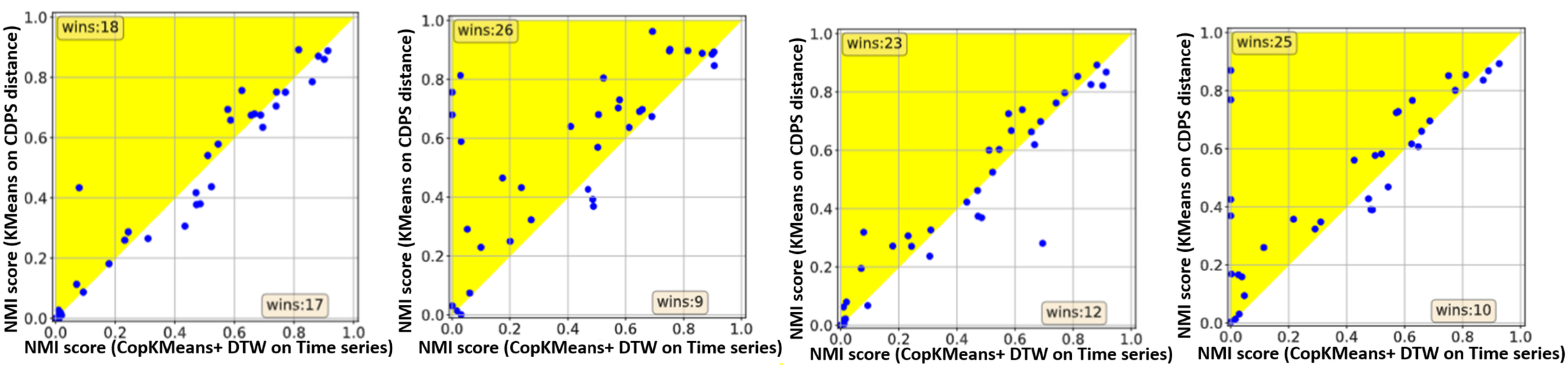
Results

Comparing clustering results using CDPS embeddings vs using raw time series on different datasets taken from the UCR archive[e]. NMI (Normalized Mutual Information) score is used to record the quality of the results, between the ground truth labels and predicted labels, a value of 1 being the best.

- **Transductive setting:** The whole dataset is used for training and testing.
- **Inductive setting:** The dataset is split into train and test sets. Constraints are introduced in training.
- **CopKMeans:** Constrained version of KMeans – constraints are provided before clustering and monitored for violation during the clustering, if so, no solution is found.

[e]https://www.cs.ucr.edu/~eamonn/time_series_data_2018/





(a) Transductive Setting. Left no constraints, Right 25% constraints

(b) Inductive Setting. Left no constraints, Right 25% constraints

Figure 2: Clustering Results.

Conclusion

- We extended LDPS to a semi-supervised setting called CDPS. We reported an increase in clustering quality compared to LDPS.
- CDPS results in a Euclidean space therefore lays the groundwork to measure constraint properties (Informativeness/Coherence, see above).
- This paves the way for constraint proposition (proposed by the algorithm to the expert) to be tackled as future work.