Time Series Clustering based on DTW and Soft-DTW

MERLIN ROUDIER, EMILIE POISSON CAILLAULT
Université du Littoral Côte d'Opale, Laboratoire d'Informatique Signal Image Côte d'Opale

Objective

Missing data appear in almost domains of applied sciences. When analyzing time series that contains missing data may lead to a loss of system efficiency and unreliable results.
To estimate missing values with the hope of high accuracy, this study proposes to use the PAM clustering algorithm using DTW and Soft-DTW distance measures to find the data characteristics of the enriched dataset generated by TimeGAN.

Methodology

To implement the methodology, we use the following algorithms and functions:
TimeGAN to generate large numbers of points and observations using a Generative Adversarial Network [3].

GlobalFeatures (gF): The function retrieves information for comparison when searching for viable windows: Minimum; Maximum; Average; Median; Standard deviation; Asymmetry, Number of peaks, Size; Entropy [2].

Dynamic Time Warping (DTW), It calculates the similarity between two time-varying signals, taking into account a single path.

Soft Dynamic Time Warping (Soft-DTW). The principle is the same as the DTW algorithm, but we take several paths into account when calculating the end cost [1].

Partitioning Around Medoids (PAM) is a clustering algorithm that groups similar data points in relation to equal central points that are always real (unlike K-means).

Conclusion & perspective

In this study, we have studied several algorithms such as TimeGAN, gF, DTW, Soft-DTW and conducted some experiments:

- **TimeGan** increases the dataset from 400 points to over 20,000 points.
- **The search for feasible windows** generally yields more than 200, providing a variety of clustering options.
- DTW and Soft-DTW matrix **Similarity** perform robustness for the clustering step.
- **Clustering** is able to identify different clusters that could be interesting depending on the criteria used.
- **Further work**: Forecasting by CNN-GRU network architecture.

Overview of the model

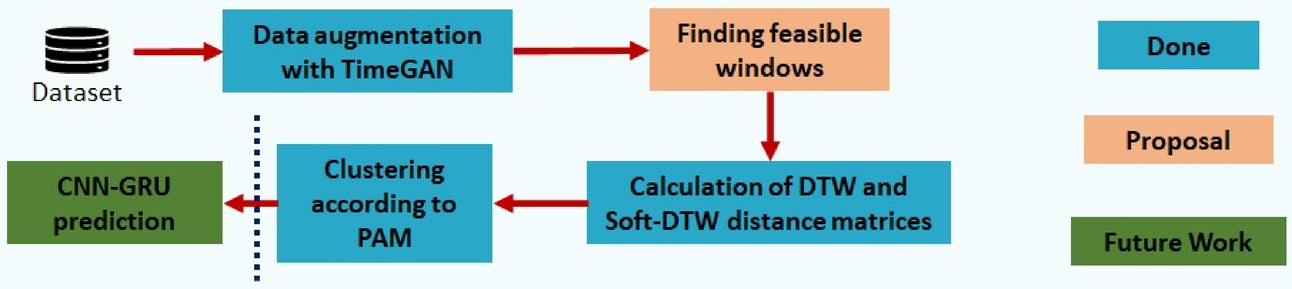

Fig. 1: Proposed diagram based on HTSCG [4]

Method for finding feasible windows

1. Define the size of the reference window and considered window.
2. Extract global Features to obtain the various information for the reference and each considered windows.
3. Compute cosine measure between gF of reference and considered window. The considered window will be keep if this score is greater than 0.995.

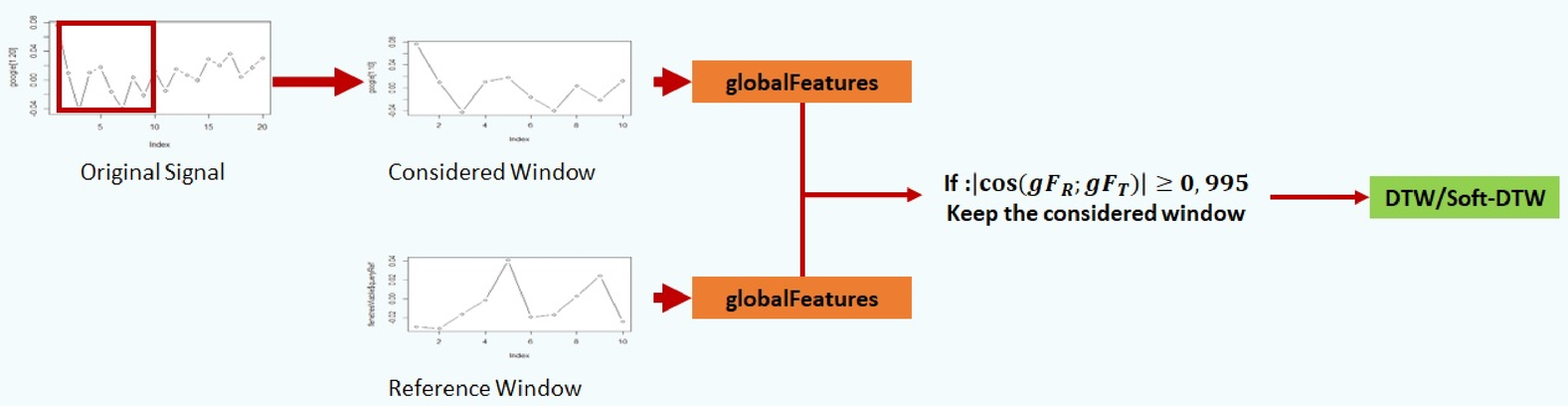

Fig. 2: Diagram of feasible window selection

PAM clustering results

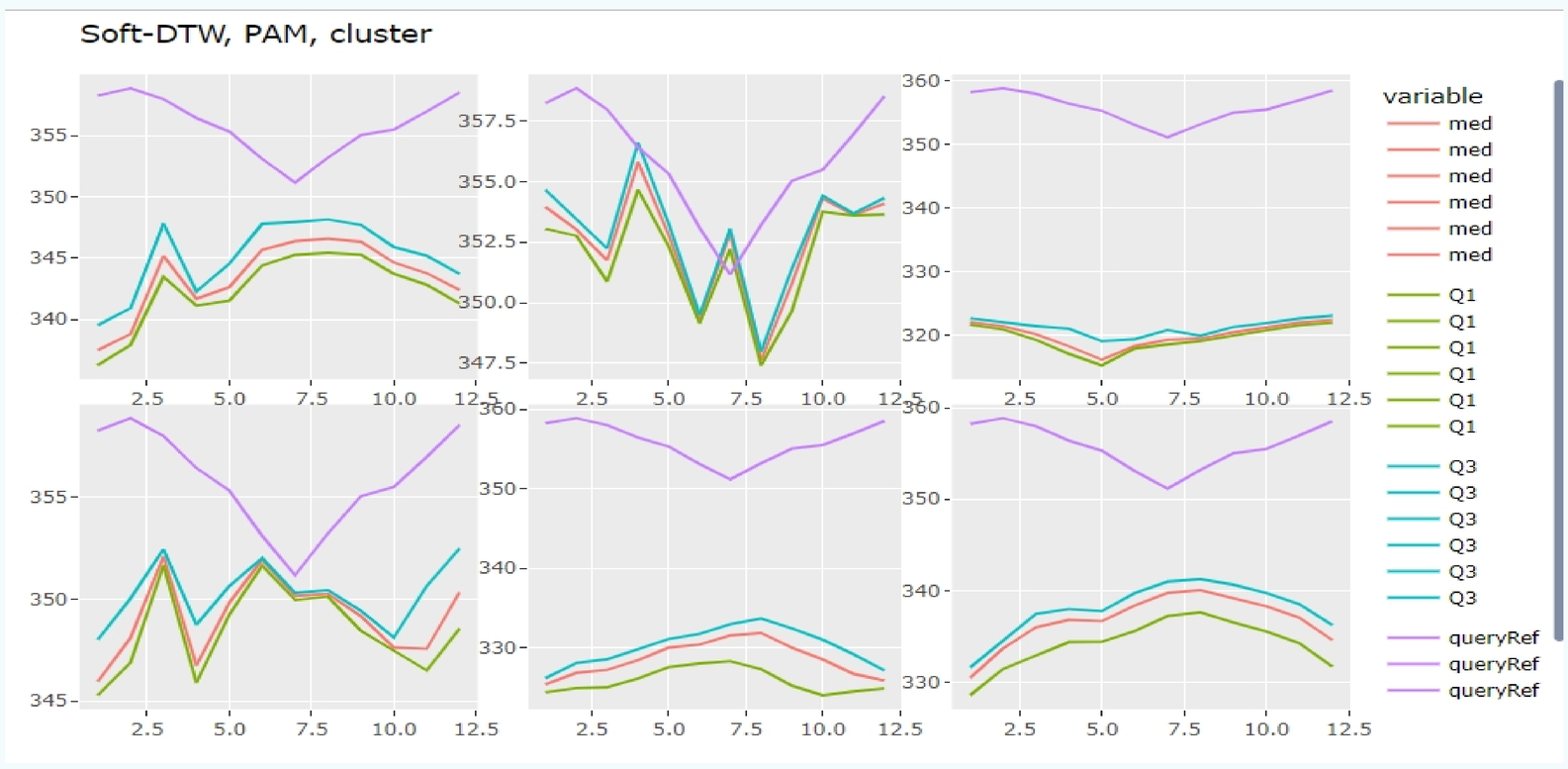

Fig. 3: Clusters graph on R co2 dataset

There are two competing criteria for selecting the cluster to be taken:

1. Find the lowest average DTW/Soft-DTW cost among all clusters.
2. Level all median curves in each cluster at the last point before the gap. Calculate the cost dtw of each in relation to the reference Query and take the cluster with the lowest cost.

References

[1] Cuturi, Marco & Blondel, Mathieu. (2017).*Soft-DTW: a Differentiable Loss Function for Time-Series.*
[2] Phan, Thi-Thu-Hong. (2018). *Elastic matching for classification and modelisation of incomplete time series.*
[3] Yoon, Jinsung & Jarrett, Daniel & Schaar, Mihaela. (2019). *Time-series Generative Adversarial Networks*.
[4] Li, Qing & Zhang, Xinyan & Ma, Tianjiao & Liu, Dagui & Wang, Heng & Hu, Wei. (2022).*A Multi-step ahead photovoltaic power forecasting model based on TimeGAN, Soft DTW-based K-medoids clustering, and a CNN-GRU hybrid neural network. Energy Reports*.8. 10346-10362. 10.1016/j.egyr.2022.08.180.

Partners