Exploring Interpretable Features for Large Time Series with SE4TeC

Jingwei ZUO, Karine ZEITOUNI and Yehia TAHER
University of Versailles Saint-Quentin-en-Yvelines, University of Paris-Saclay, Versailles, France
{firstname.lastname}@uvsq.fr

Background

Shapelet[1]: A representative shape in Time Series which is capable of distinguishing one class from the others

Time Series(TS) Applications:

- Medical diagnosis
- Industrial troubleshooting
- Human Activity Recognition
- Astronomy Discovery, etc.

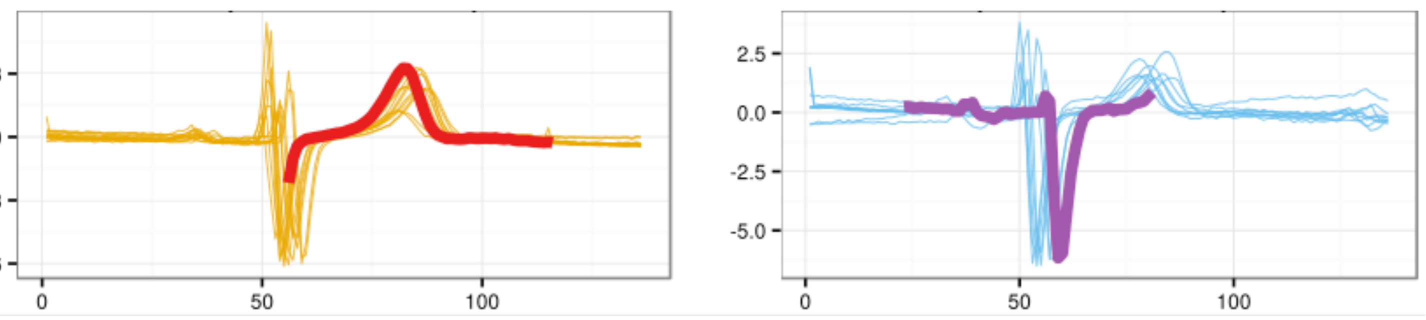

Figure 1: Two classes from the "ECGFiveDays" dataset and the *best representative patterns (Shapelets)*

Problem Statement

State-of-the-art Shapelets Extraction algorithms:

- **High** computation cost
- **Low** scalability in Big Data context
- **Low** interpretability during extraction process

Proposals and System structure

Main idea of SE4TeC (Scalable Engine For efficient and expressive Time series Classification) :

- Computations should be shared and executed independently
- Small communication cost between the distributed nodes
- Evaluation of the importance of candidate Shapelet could be done partly at local and then be conducted globally

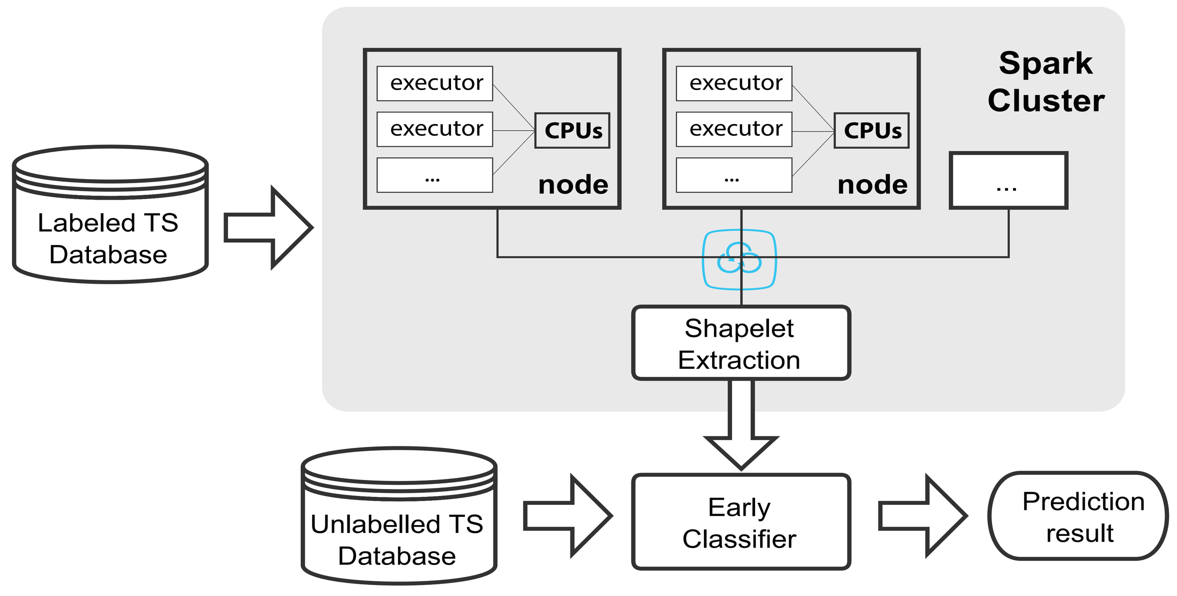

Figure 2: Distributed structure of SE4TeC

Interpretability

Distance Profile[2]: A vector which stores the distance between a given subsequence/query $T_{i,m}$ and every subsequences $T'_{j,m}$ of a target Time Series T'. Formally, $DP^m_{i,j} = dist(T_{i,m}, T'_{j,m}), \forall j \in [0, n'-m+1]$

Matrix Profile: A vector of distance between subsequence $T_{i,m}$ in source T and its nearest neighbor $T'_{j,m}$ in target T'. Formally, $MP^m_i = min(DP^m_i)$, where $i \in [0, n-m+1]$

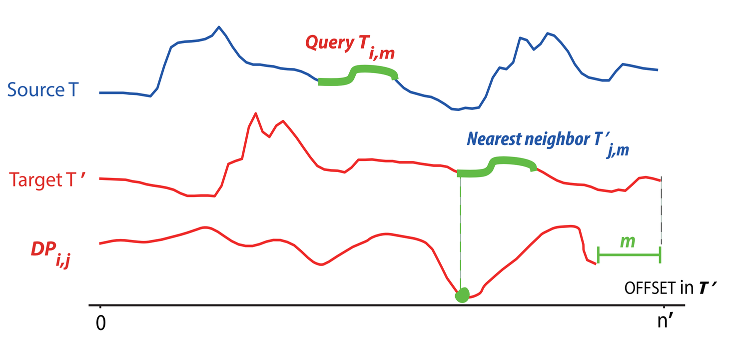

Figure 3: *Distance Profile between Query $T_{i,m}$ and target time series T', where n' is the length of T'. $DP_{i,j}$ can be considered as a meta TS annotating target T'*

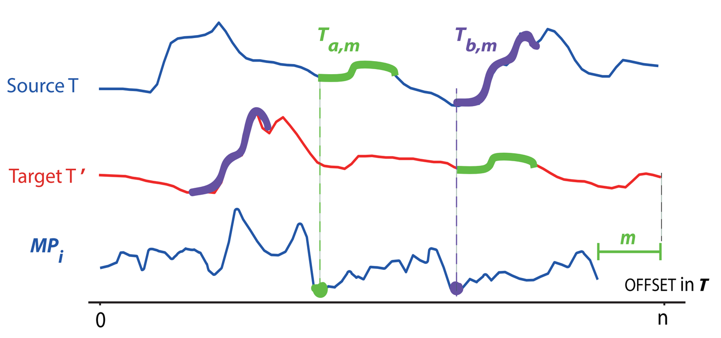

Figure 4: *Matrix Profile between Source T and Target T', where n is the length of T. Intuitively, MP_i shares the same offset as source T*

Shapelet on MAtrix Profile (SMAP)

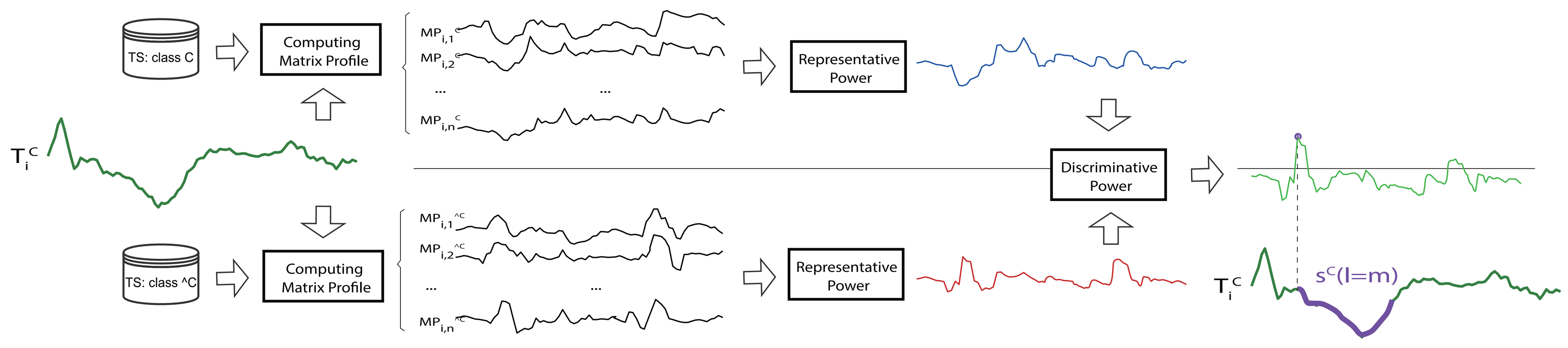

Figure 5: Interpretable Extraction process of Shapelet

- **Assessment of Shapelets in batches**
  1. **Representative Profile:**
     - Shows a vector of representative power of the subsequences in a Time Series
     - For each subsequence, compute its average distance to the instances in dataset:

$$RP(T_i^C, D) = avg(MP_{T_i^C, T_j})$$

  2. **Discrimination Profile:**
     - For each subsequence, compute the difference of Representative Power from class C to others (OVA, one-vs-all):

$$Discm_{Profile}(T_i^C, D) = -(RP(T_i^C, D^C) - RP(T_i^C, D^{!C}))$$

  Advantages of $Discm_{Profile}$:
  - A split distance can be given directly to check the inclusion between Shapelet and Time Series
  - $Discm_{Profile}$ VS *Information Gain* in time: $\mathcal{O}(1)$ | $\mathcal{O}(N^2n^2)$
    (N is the number of TS, n is the length of the longest TS in dataset)

- **Inclusion between Shapelet & TS:**

$$sInT(T, \hat{s}^C) = \begin{cases} true, & if\ dist(T, \hat{s}^C) \leq RP(\hat{s}^C, D^C) \\ false, & otherwise \end{cases}$$

Optimization of SMAP:

- There is a big probability that the Nearest Neighbor stays on the same offset when the length of query increases:

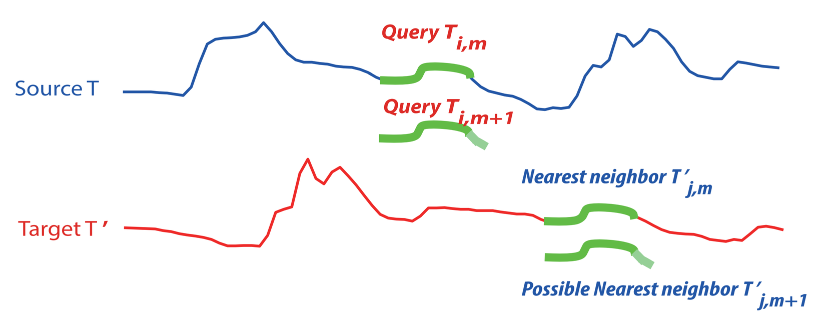

Figure 6: Optimization of SMAP by defining a Lower Bound Distance

- Lower Bound Distance[4]: Based on $dist(T_{i,l}, T_{j,l})$, we can estimate a min. possible value of $dist(T_{i,l+k}, T_{j,l+k})$ to accelerate the calculation of $\min(DP_i^m)$, other than computing the whole Distance Profile.

Experiments & Results

The program is executed on AWS EMR cluster. The baseline is **USE[3]**, which utilizes the traditional method for shapelet extraction based on Information Gain. **1NN** classier is applied for all security tests. Two real-life sensor datasets are tested on AWS EMR cluster:

- ECG medical diagnosis **(ECG200): 100** labelled records, Length=96
- Wafer industrial troubleshooting **(Wafer):1000** labelled records, Length=152

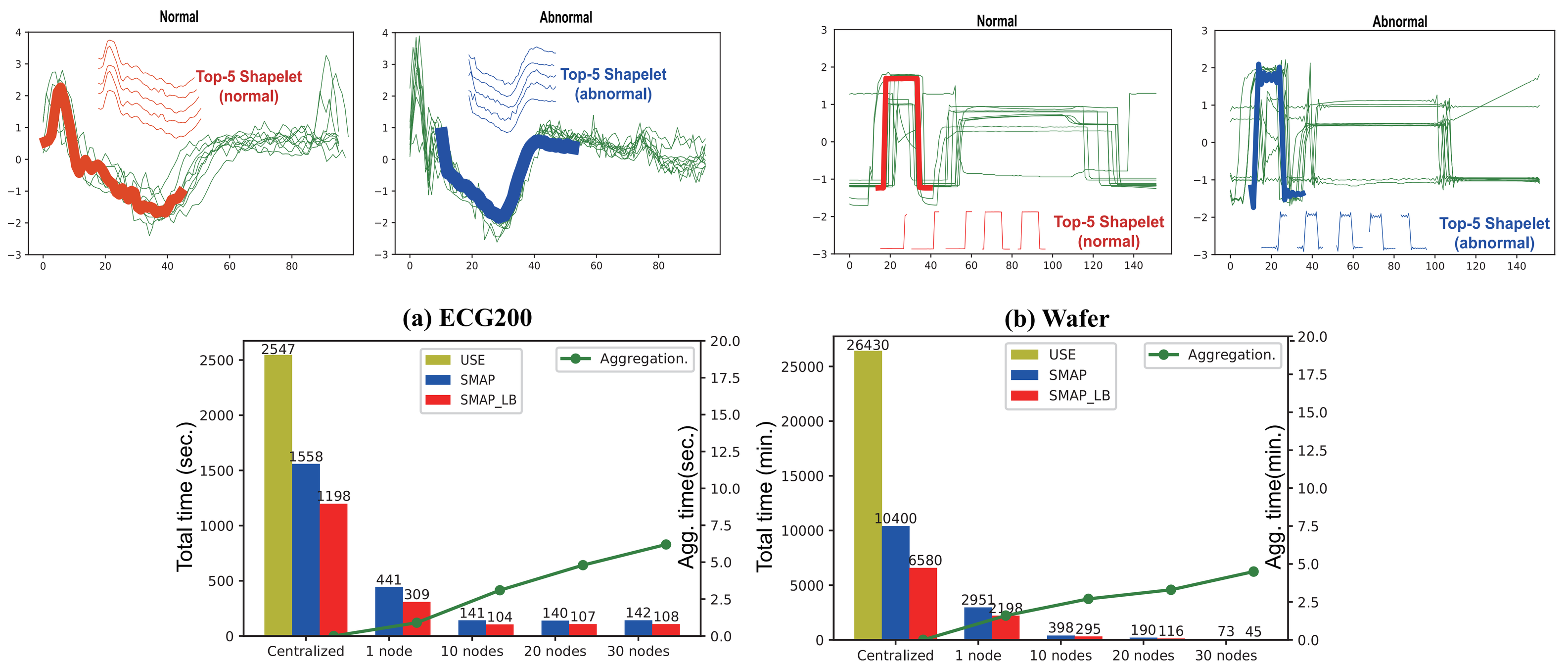

Figure 7: Shapelets extracted from datasets & Performance comparison

From 1 to 30 nodes on cluster mode:

- the total time cost drops to **0.68%**.
- the communication cost between distributed nodes for **Wafer** increases by **181%**

Considering the gain, the communication cost can be ignored when we expand the cluster to a larger scale. More details and application results of SE4TeC can be found in our project page: https://github.com/JingweiZuo/SE4TeC

Reference

1. *Lexiang Ye and Eamonn Keogh. Time series shapelets: A New Primitive for Data Mining. In Proc. SIGKDD 2009*
2. *Chin-Chia Michael Yeh et al. Matrix Pro le I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. In Proc. ICDM 2016*
3. *Raef Mousheimish, Yehia Taher, and Karine Zeitouni. Automatic Learning of Predictive CEP Rules: Bridging the Gap between Data Mining and Complex Event Processing. In Proc. DEBS '17. 158–169*
4. *Michele Linardi et al. VALMOD: A Suite for Easy and Exact Detection of Variable Length Motifs in Data Series. In Proc. SIGMOD'18.*