ISET: Incremental Shapelet Extraction from Time Series Stream

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

Background

Time Series Representations	
R1	Global features of entire series (1-NN)
R2	Summary statistics of sub-series
R3	Motif (frequent patterns)
R4	***Shapelet[1] (shape-based features)***

Data Stream Challenges	
C1	***Infinite Length***
C2	***Feature Evolution***
C3	***Concept Drift***
C4	Concept Evolution

Streaming Time Series S

- A continuous input data stream where each instance is a real-valued data: $S = (t_1, t_2, \dots, t_N)$, where N is the time tick of the most recent input value.

Time Series Stream S_{TS}

- A continuous input data stream where each instance is a Time Series: $S_{TS} = (T_1, T_2, \dots, T_N)$. Notice that N increases with each new time-tick.

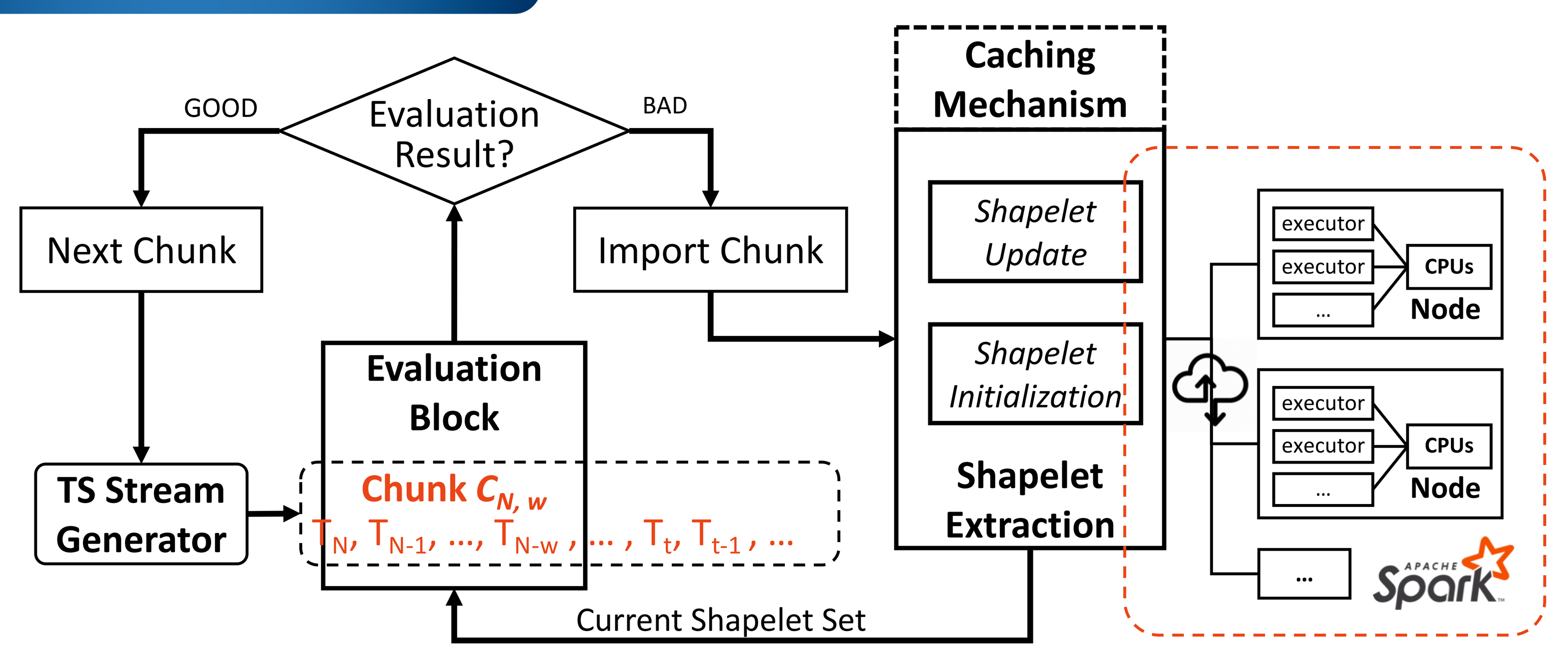
Research Focus:

- R4 + {C1, C2, C3} in Time Series Stream.

Problem Statement

- Low Scalability and Incrementality of Time Series representation approaches
- Classical Shapelet Evaluation is not suitable in streaming context
- Concept Drift detection should be adapted in TS Stream model
- Memory cost of infinite TS instances (Shapelet Extraction relies on a set of instances cached in the memory)

System Structure



Scalability & Incrementality

Scalability:

Previous work[2] ensures the scalability of Shapelet Extraction in Spark.

Incrementality:

- *The necessary condition* to adapt TS representation in stream context.
- When new TS instance comes:
  1. Update the discriminative power of existing Shapelets
  2. Introduce new candidate Shapelets, compute their power
- Step 1 and 2 share the same computation process[4]

Evaluation Block (Shapelet Evaluation + Concept Drift Detection)

Shapelet Evaluation

- 0-1 Loss Function

$$L(Y, h(T)) = \begin{cases} 0, & Y = h(T) \\ 1, & Y \neq h(T) \end{cases}$$

where

$$h(T) = \begin{cases} C, & if\ dist(T, \hat{s}) \leq \hat{s}.dist_{Thresh} \\ nonC, & otherwise \end{cases}$$

- Sigmoid Loss Function

$$L(Y, h(T)) = \frac{1}{1 + e^{-(x-\sigma)}}, \quad \sigma = \hat{s}.dist_{Thresh}$$

$$x = min(dist(T^C, \hat{s})), \quad \hat{s} \in \hat{S}^C$$

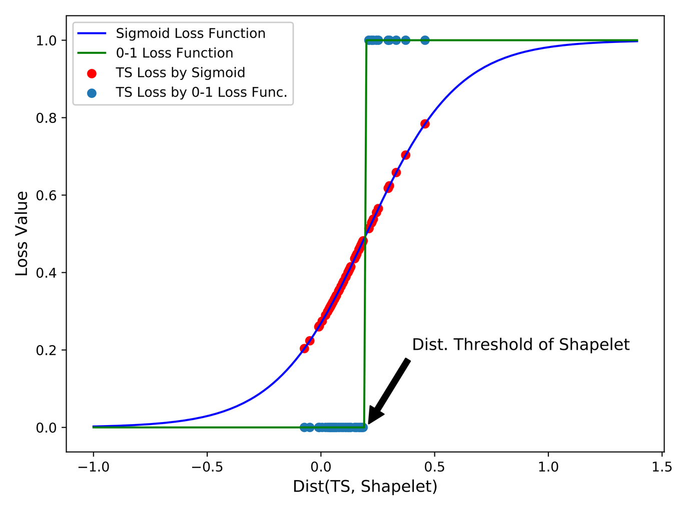


Figure 1: Shapelet Evaluation over newly input TS instances

A Loss Threshold Δ can be set to import incrementally the valuable instances.

Evaluation Block (cont.)

Concept Drift detection

- Page-Hinkey (PH) Test[3]: a typical technique for change detection in signal processing.

$$L_C(N) = \frac{1}{w} \sum_{k=1}^{w} L(Y_{N-w+k}, h(T_{N-w+k}))$$

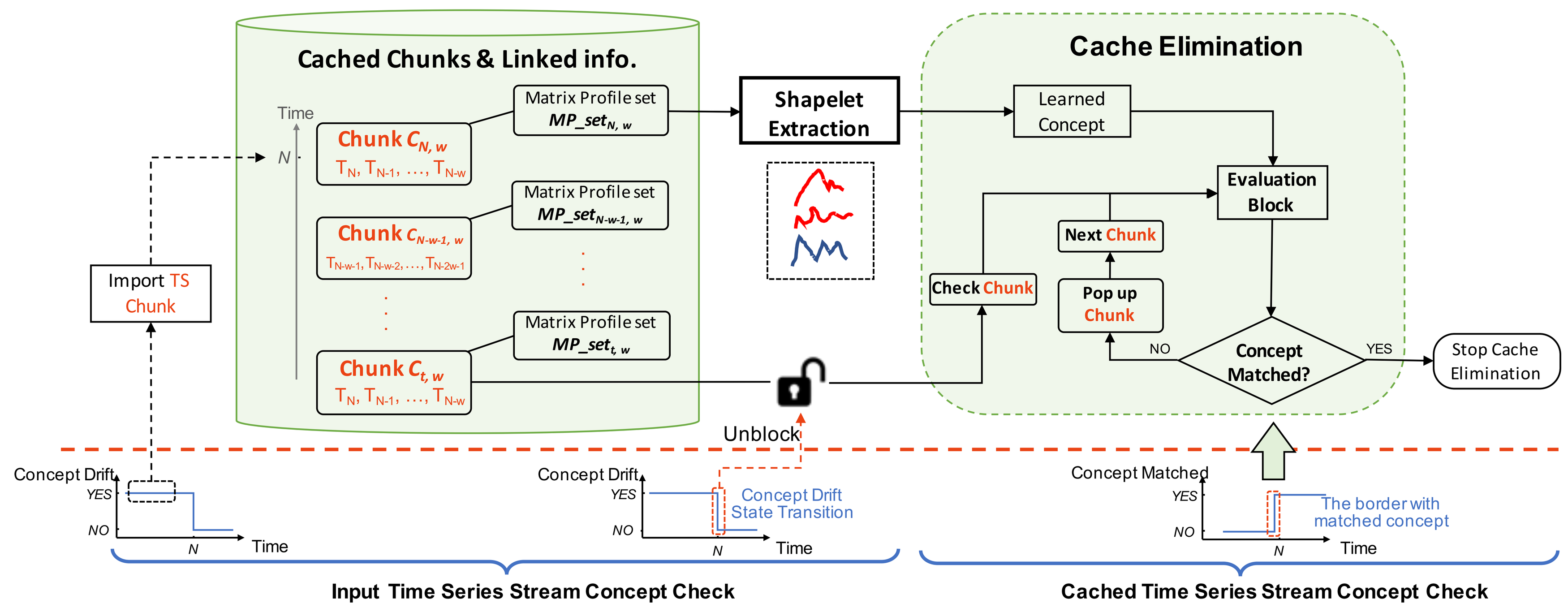
$$m_N = \sum_{t=0}^{N} (L_C(t) - L_{avg}(t) - \delta)$$

$$M_N = min(m_t, t = 1 \dots N)$$

$$PH_N = m_N - M_N$$

- $L_C(N)$: the average loss of newly input TS chunk
- m_N: the cumulative difference between the chunk loss and average loss until the current time. δ: Loss Tolerance
- M_N: the minimal cumulative difference recorded
- λ: PH threshold to detect a Concept Drift
- $Concept\ Drift = \begin{cases} True, & PH_N \geq \lambda \\ False, & otherwise \end{cases}$

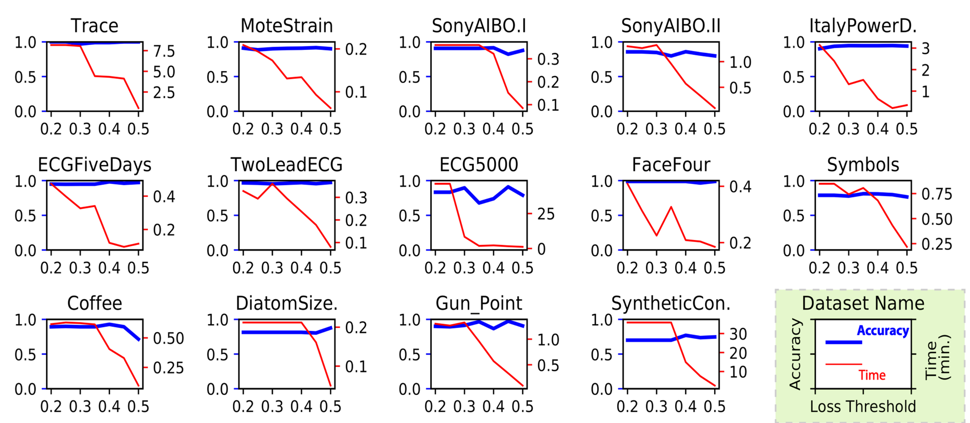
Elastic Caching Mechanism



Experimental Results

- Incremental test under stable concept (14 Shapelet datasets)

Type	Name	Train/Test	Class	Length	IG	KW	MM	ISMAP(best)	Para. (Δ)	Comp. Ratio
Simulated	SyntheticControl	300/300	6	60	**0.9433**	0.9000	0.8133	0.7007	0.35	46.7%
Sensor	Trace	100/100	4	275	0.9800	0.9400	0.9200	**1**	0.5, 0.45	26.0%
	MoteStrain	20/1252	2	84	0.8251	0.8395	0.8395	**0.9169**	0.45	60.0%
	SonyAIBO.I	20/601	2	70	0.8453	0.7281	0.7521	**0.9151**	0.4	95.0%
	SonyAIBO.II	27/953	2	65	0.8457	-	-	**0.8583**	0.4	63.0%
	ItalyPower.	67/1029	2	24	0.8921	0.9096	0.8678	**0.9466**	0.45	25.4%
ECG	ECG5000	500/4500	5	140	0.7852	-	-	**0.9109**	0.4	9.4%
	ECGFiveDays	23/861	2	136	0.7747	0.8721	0.8432	**0.9826**	0.4	51.2%
	TwoLeadECG	23/1189	2	82	0.8507	0.7538	7657	**0.9337**	0.5	47.8%
Images	Symbols	25/995	6	398	0.7799	0.5568	0.5799	**0.8113**	0.35	96.0%
	Coffee	28/28	2	286	**0.9643**	0.8571	0.8671	0.9286	0.4	78.6%
	FaceFour	24/88	4	350	0.8409	0.4432	0.4205	**0.9886**	except 0.45	62.5%
	DiatomSize.	16/306	4	345	0.7222	0.6111	0.4608	**0.8758**	0.5	50.0%
Motion	GunPoint	50/150	2	150	0.8933	0.9400	0.9000	**0.9733**	0.45	42.0%



Accuracy Performance

Baseline: Shapelet Tree classifiers

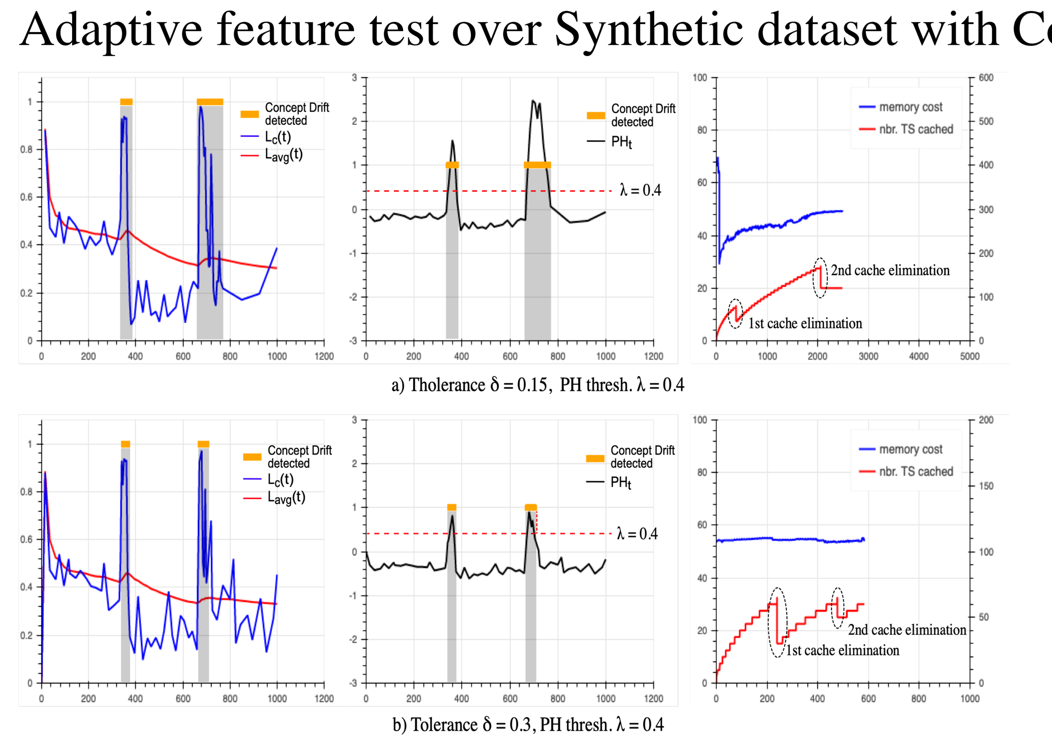
- Information Gain (IG)[1]
- Kruskall-Wallis (KW)[5]
- Mood's Median (MM)[5]

$$Comp.Ratio = \frac{nbr.instance_{imported}}{nbr.instance_{training}}$$

Trade-off between Accu. and Δ

- **In theory**, the higher the loss threshold Δ, the higher the efficiency, the lower the accuracy
- **In practice**, the highest accuracy falls in the range $\Delta \in [0.35, 0.45]$. Nevertheless, efficiency can be greatly increased with an exchange of a negligible decrease of accuracy.

- Adaptive feature test over Synthetic dataset with Concept Drift



Synthetic *Trace* dataset:

- Randomly put noise for Data Augmentation
- 1000/1000 training/testing instances
- Two drifts are inserted at time *333* and *667*

Concept Drift detection:

- 345/330 (δ=0.15), 350/330 (δ=0.30)

Caching cost:

- 100 of 1000 (δ=0.15), 50 of 1000 (δ=0.30)
- Cache is eliminated at the end of drift period

TABLE I: Reliability of Extracted Shapelets on 4 time ticks at the beginning/end of each drift area

Dataset	-	**i**(Con. 1)	**ii**(Con. 2)	**iii**(Con. 2)	**iv**(Con. 3)
$Aug.Trace(\delta = 0.15)$	Time tick	*345*	*380*	*670*	*790*
	Test Accu.	0.9600	0.9900	0.9900	0.9800
$Aug.Trace(\delta = 0.30)$	Time tick	*350*	*365*	*675*	*700*
	Test Accu.	0.9600	0.9800	0.9800	0.9700

Conclusion

- ✓ First attempt to explore incremental and adaptive features in Time Series Stream.
- ✓ We propose a novel Shapelet Evaluation approach which allows the transition from Time Series to Data Stream analysis.
- ✓ We propose an elastic caching mechanism which is capable of eliminating out-of-date concepts/data proactively in the Time Series Stream model.
- ✓ The system is applicable in the scenario where an existing dataset is continuously expanded with new knowledge without human loop in the middle.

References

1. Lexiang Ye and Eamonn Keogh. "*Time series shapelets: A New Primitive for Data Mining*." In Proc. SIGKDD 2009
2. J. Zuo, K. Zeitouni, and Y. Taher, "*Exploring interpretable features for large time series with SE4TeC*." In: EDBT 2019, Lisbon, Portugal. pp. 606–609 (2019)
3. J. Gama, I. Zliobait E, A. Bifet, M. Pechenizkiy, and A. Bouchachia. "*A Survey on Concept Drift Adaptation*." ACM Comput. Surv. 1, 1, Article, vol. 1, 2013.
4. J. Zuo, K. Zeitouni, and Y. Taher, "*Incremental and Adaptive Feature Exploration over Time Series Stream*", AALTD@ECML-PKDD'19
5. Jason Lines, and Anthony Bagnall, "*Alternative Quality Measures for Time Series Shapelets*", IDEAL 2012

Project Page in Github

ECML-PKDD'2019 – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Würzburg, Germany.