ISETS: Incremental Shapelet Extraction from Streaming Time Series

Jingwei Zuo[1,2], Karine Zeitouni[1,2], and Yehia Taher[1,2]

[1] DAVID Lab. University of Versailles, France
[2] University of Paris-Saclay, France
{jingwei.zuo, karine.zeitouni, yehia.taher}@uvsq.fr

Abstract. In recent years, Time Series (TS) analysis has attracted widespread attention in the community of Data Mining due to its special data format and broad application scenarios. An important aspect in TS analysis is Time Series Classification (TSC), which has been applied in medical diagnosis, human activity recognition, industrial troubleshooting, etc. Typically, all TSC work trains a stable model from an off-line TS dataset, without considering potential Concept Drift in streaming context. Conventional data stream is considered as independent examples (e.g., row data) coming in real-time, but rarely considers real-valued data coming in a sequential order, called Streaming Time Series. Processing such type of data, requires combining techniques in both communities of Time Series (TS) and Data Streams. To facilitate the users' understanding of this combination, we propose *ISETS*, a web-based application which allows users to monitor the evolution of interpretable features in Streaming Time Series.

1 Introduction

Time Series (*TS*) is a sequence of real-valued data, which can be collected from various sources, such as ECG data in medicine, IoT data in smart cities, light-curves in astronomy, etc. In this work, we study the problem of Streaming Time Series Classification (*STSC*): given a Streaming *TS* source, we aim at learning incrementally the concept allowing to predict the class of new input *TS* unit, and catching the concept drift in the data flow.

Concept [1] refers to the target variable, which the learning model is trying to predict. Existing work in data streams is mostly based on the assumption that data instances are independently and identically distributed (i.i.d) within a particular concept. Most *TSC* approaches are biased towards learning a stable concept from an off-line Time Series dataset, but not adaptable to streaming concept-drifting context, where a gradual change of the concept happens along with the input of TS streams. Lazy classifiers such as Nearest Neighbor (1-NN) [5] and dictionary based approaches [4] are applicable for *STSC*. However, every input instance will be considered to adjust the inner concept, which requires potentially a large buffer space and will bring a huge computation cost.

Our proposal, namely ISETS: **I**ncremental **S**hapelet **E**xtraction from Streaming **T**ime **S**eries, is capable of building the gap between Time Series Classification and Data Streams processing. Based on Shapelets [6], interpretable shapes considered as features in Time Series, the web-based application allows users to capture an adaptive concept for new incoming TS with a small memory buffer and a minimal computation cost. Besides, ISETS possesses a highlighted interpretability, as well as scalability in Big Data context. All implementation code, testing datasets and video tutorial are available online[3].

2 System Structure

As shown in **Fig.** 1, the system is composed by two blocks, namely Shapelet Extraction and Concept Drift Detection. By applying recent extracted Shapelets on new incoming TS streams, we can decide whether or not to cache TS instances into memory according to the Concept Drift Detection. From newly cached TS instances, Shapelet Extraction Block will make use of historical computations and update Shapelet Ranking at a minimal cost.

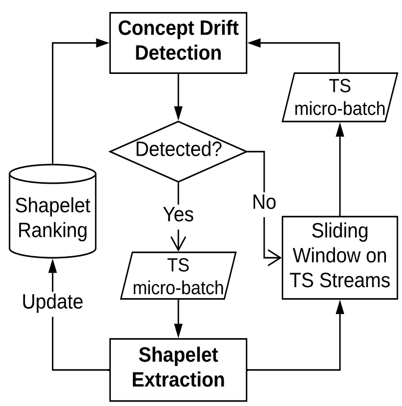

Fig. 1: Main system structure

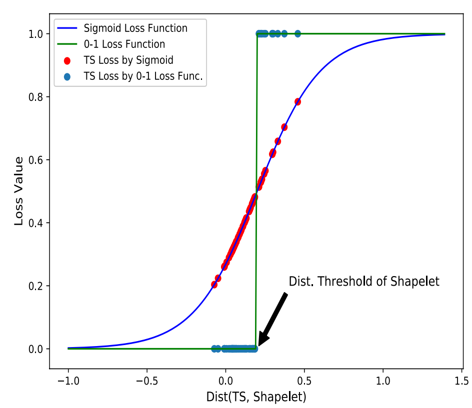

Fig. 2: Loss measure of Time Series by Sigmoid Function and 0-1 Loss Function

Concept Drift Detection: To detect the concept drift, a simple test can be done by comparing the average loss of current catching TS micro-batch and that of all historical TS. Page-Hinkley test [3] is applied here as well for an advanced detection. The classical Shapelet-based approach [6] supposes that a Time Series T can be classified by the inclusion of a class-specified Shapelet $\hat{s}$. (i.e. if $dist(T, \hat{s}) < \hat{s}.dist_{thresh}$, then $T.class = \hat{s}.class$). However, two Time Series with similar distance to a Shapelet may obtains different classes by this strategy. A loss measured by a crisp *0-1 Loss Function* is then ill-adapted. To this end, we propose a loss measure based on Sigmoid Function, to convert the inclusion problem to the possibility that a TS contains the Shapelet. The

[3] https://github.com/JingweiZuo/ISETS

loss distribution is shown in **Fig.** 2. Every loss under 0.5 represents a relative acceptable classification result. Intuitively, the cumulative loss represents the adaptability of extracted Shapelets to the current TS micro-batch. Moreover, a forgetting mechanism is proposed when the most recent data are deemed more important. To this end, we apply an exponential moving sum to the loss.

Incremental Shapelet Extraction: The Shapelet Extraction is based on SE4TeC proposed in [7], but with the consideration of streaming data context. A set of Shapelets of different classes will be updated once a Concept Drift is detected, which means only the Time Series beyond the current concept will be taken into account by the computation. Each Shapelet will be given a score for its discriminative power between the classes.

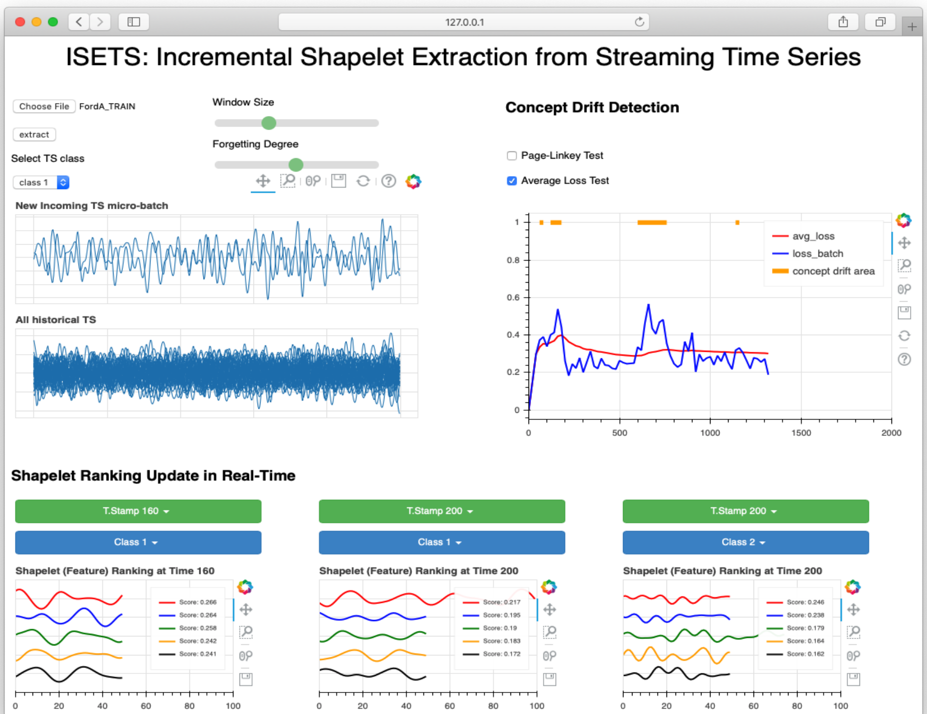

Fig. 3: GUI of ISETS web application

3 About the Demonstration

Through this demonstration, attendees will have the opportunity to explore interpretable Shapelet features and Concept Drift in the context of Streaming Time Series. A web application with GUI shown in **Fig.** 3 allows an interactive use of the system. For the operations, users can adjust the sliding Window to set the size of input TS micro-batch. By changing system's forgetting degree, users can control the importance of recent coming data on current concept. As the result, our system allows monitoring the occurrence of Concept Drift and the evolution of Shapelet Ranking of each class at different time points. We show in **Fig.** 3 the intermediate results of the test on FordA dataset [2], which contains 3601 labelled Time Series with a fixed length of 500. The concept drift time periods are marked, where the new incoming TS micro-batches are considered by Shapelet Extraction Block to update the Shapelet Ranking. We can easily capture the Shapelets from different classes and time points.

The Shapelet Extraction process can be either conducted at local or on a remote Spark cluster. We provide also an 1-click cluster based on Docker, to facilitate the replay of the distributed test offline by the user[4].

4 Conclusion

In this paper, we present a novel approach, namely ISETS, to bridge the gap between Time Series Classification and Data Streams analysis. A web application is provided to facilitate attendees to interact with the system. ISETS allows users to detect the Concept Drift within Streaming Time Series, and monitor the evolution of TS features (i.e., Shapelet) in an interpretable way.

Acknowledgements

This research was supported by DATAIA convergence institute as part of the *Programme d'Investissement d'Avenir*, (ANR-17-CONV-0003) operated by DAVID Lab, University of Versailles Saint-Quentin, and MASTER project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie-Slodowska Curie grant agreement N. 777695.

References

1. Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B.: MOA: Massive Online Analysis. Tech. rep. (2010)
2. Dau, H.A., Bagnall, A., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Keogh, E.: The ucr time series classification archive (2018), `https://arxiv.org/pdf/1810.07758.pdf`
3. Gama, J., Zliobait E, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A Survey on Concept Drift Adaptation. ACM Comput. Surv. 1, 1, Article **1** (2013)
4. Lin, J., Khade, R., Li, Y.: Rotation-invariant similarity in time series using bag-of-patterns representation. J INTELL INF SYST **39**(2), 287–315 (2012)
5. Ueno, K., Xi, A., Keogh, E., Lee, D.J.: Anytime classification using the nearest neighbor algorithm with applications to stream mining. Proc. ICDM'06 (2006)
6. Ye, L., Keogh, E.: Time series shapelets: A New Primitive for Data Mining. Pro. KDD'09 p. 947 (2009)
7. Zuo, J., Zeitouni, K., Taher, Y.: Exploring Interpretable Features for Large Time Series with SE4TeC. Proc. EDBT 2019 pp. 606–609 (2019)

[4] https://github.com/JingweiZuo/ISETS/tree/master/Spark_Cluster_Docker