Time Series meet Data Streams: Perspectives of the Interdisciplinary Collision and Applications

Jingwei Zuo, Karine Zeitouni, Yehia Taher

Time Series meet Data Streams: Perspectives of the Interdisciplinary Collision and Applications

Jingwei Zuo
DAVID Lab, University of Versailles
Université Paris-Saclay
Versailles, France
jingwei.zuo@uvsq.fr

Karine Zeitouni
DAVID Lab, University of Versailles
Université Paris-Saclay
Versailles, France
karine.zeitouni@uvsq.fr

Yehia Taher
DAVID Lab, University of Versailles
Université Paris-Saclay
Versailles, France
yehia.taher@uvsq.fr

ABSTRACT

Time Series (TS) analysis has been always a research hotspot in Data Mining community, due to its complex sequential structure and board application scenarios. Typically, the analysis process is launched from an off-line TS dataset, without considering the context with dynamic data source. For instance, domains like healthcare look to enrich the database gradually with more medical cases, or in astronomy, with human's growing knowledge for the universe, the theoretical basis for labelling data will change. The techniques applied in a stable TS dataset are then not adaptable in such dynamic scenarios, that said streaming context. Classical data stream analysis are biased towards vector or row data, where each attribute is independent to train an adaptive learning model, but rarely considers Time Series in a stream format. Processing such type of data, requires combining techniques in both communities of Time Series (TS) and Data Streams (DS). To this end, we take the first attempt to study the gap between the two communities, and give an overview for the interdisciplinary challenges and applications.

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, GPS or accelerometer data in activity recognition, etc. Time Series Classification (TSC) is intended to predict the label of a newly input TS instance by extracting the knowledge from collected data. The optimization of TS feature extraction and model construction process allows us to strive for a low prediction error, and approach to Time Series' nature Concept [1], which refers to the target variable that the learning model is trying to predict. When the research context extends to Data Streams (DS), the knowledge base is no longer constant and evolves gradually with newly input data. The challenges here can be represented by three intrinsic characteristics of Data Stream [3]:

- **Feature Evolution:** The feature space changes over time, new features become useful and old one may become redundant. The learning model is intended to extract appropriate features to approach the inner concept of data source.
- **Concept Evolution:** Non-labelled instances with novel classes may emerge in the future, which should be recognized and separated from existing classes.
- **Concept Drift:** The prediction target evolves over time, the learning model must be capable of adjusting itself gradually to the most recent data by an effective process of Concept Drift detection.

Then, the exploration of TS features plays a key role in bridging TS and DS analysis. Various TS feature representations are applied in Time Series Classification (TSC):

R1: Global feature of entire series [4] for *1-NN* classifier
R2: The summary statistics (e.g., mean, deviation, slope, etc.) [2, 9] extracted from every sub-series for ensemble classifiers
R3: Motif [10] features when frequent patterns characterize a class
R4: Shapelet [11] when specific sub-series determines a class

Apparently, single feature representations can be combined to construct ensemble classifiers [5, 6] which lead to the state-of-the-art model accuracy. However, a lightweight feature representation is preferred to eliminate the model collision from TS to DS context. In this paper, we will give answers for the following questions:

Q1: How to define a *Data Stream (DS)* in the context of Time Series?
Q2: Under different *DS* definitions, what are the collision and cooperative applications between the community of TS and DS?

Then we will present briefly our preliminary work covering some of *TS-DS* scenarios. Future work is planned in the conclusion.

2 STREAM DEFINITION IN TS CONTEXT

Definition 1: *Time Series Stream* S_{TS} is a continuous input data stream where each instance is a Time Series: S_{TS}=(T_1, T_2, ..., T_N). Notice that N increases with each new time-tick.

The information should be extracted from each newly input TS instance, and be merged with existing learning model. As shown in *Table* 1, by considering various challenges in Data Streams, we are capable of launching the analysis from different contexts. Within a stationary concept, the learning model is trying to make the learned concept stay as close as the real one. Feature Evolution concerns the incrementality of learning algorithm, which is the necessary condition for stream learning system. Further consideration of non-stationary concept like Concept Evolution and Concept Drift can be respectively adopted in more dynamic contexts where data instances are weakly labelled or prediction target evolves over time.

Table 1: Various contexts in TS Stream analysis

	Feature Evolution	Concept Evolution	Concept Drift
Context 1	✓	✗	✗
Context 2	✓	✓	✗
Context 3	✓	✗	✓
Context 4	✓	✓	✓

Definition 2: *Streaming Time Series S* is a continuous input data stream where each instance is a real-valued data: S=(t_1, t_2, ..., t_i, ..., t_N), where N is the time tick of the most recent input value.

Time Series can be considered as a local collection of real-valued data from an online streaming source generating a never-ending data flow. In typical applications like patient's ECG monitoring, or analysis of data received from urban sensors, a learning model can

be built on existing labelled data set to monitor the data flow coming in real-time, that said *Training off-line*, *Monitoring on-line*. A constant feature space with stationary concept makes the processing focus on the real-time reaction [8] on input data, system efficiency becomes then the key factor in this scenario. A typical example is activity recognition [7], where the learned activity patterns serve to detect the similar ones over Streaming TS.

A further research problem comes when considering novel motif discovery from Streaming TS. The motif here represents the subseries which appears frequently over Streaming TS, the subs-series may show a regular event in data flow but unknown by the system. Motif discovery in this context requires an elastic caching mechanism for conserving information in handled data and an online monitoring process for updating sub-series' frequency. Finally, by adopting the concept *Active Learning*, the two processes can be executed in parallel with human in the loop:

1. Motif Discovery ➙*Ask for labels* ➙Add motif into Pattern Base
2. Labelled Pattern Base ➙Pattern Detection over Streaming TS

3 PRELIMINARY WORK

We start from the context of Time Series Stream, which requires basically an incremental learning process with dynamic features. Currently, we are interested in exploring interpretable TS features with an explainable extraction process, where features' evolution should be trackable over time to fit the streaming context.

Shapelet [11], as a shape-based feature in TS, which is widely adopted by the community for its reliability and interpretability, provides a possibility to fulfil the aforementioned requirements. With an advanced work in [12], the explainability of Shapelet Extraction process is ensured even to non-expert, that offers us a starting point to further explore the TS Stream context.

Here we consider the context 3 defined previously in TS Stream where Concept Drift happens over time. We are intended to extract the gradual varied Shapelets which allows an adaptive learning model to the most recent data. The system shown in **Fig.** 1 is composed by Shapelet Extraction Block and Evaluation Block. We take TS Chunk $C_{t,w}$ as minimum input unit which contains a number of continuous TS instances: $C_{t,w}$=(T_{t-w+1}, T_{t-w+2}, ..., T_t), where t is the time-tick, w is the window size.

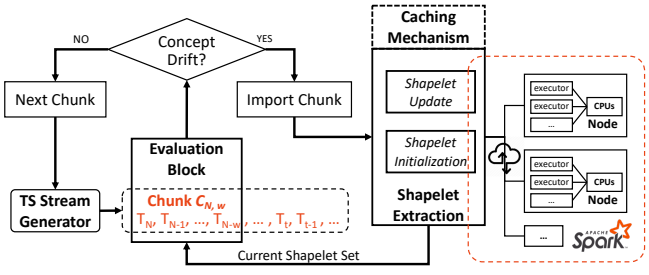

Figure 1: System Structure in TS Stream context 3

The main idea is to evaluate continuously the shift between learned concept and real concept in data source. Once a Concept Drift is detected, the input chunk will be imported into Shapelet Extraction bloc to update the learning model. Both Shapelet initialization and updating process are parallelizable on Spark cluster, which makes use of RAM as caching unit to lower the I/O cost. Besides, a caching mechanism is required by the system. As mentioned in [12], the selection of candidate Shapelets depends on their discriminative power in dataset, the fact that TS instances should be cached in memory is then a necessary condition for Shapelet Extraction. Which is the main difference compared to Concept Drift detection in classical Data Stream analysis, where it's possible to have one single pass on input instances.

4 CONCLUSION AND FUTURE WORK

In this paper, we give an overview for the combined context of Time Series and Data Streams. The specific format of Time Series brings a huge challenge on the analysis in streaming context, in particular, the adaptive feature exploration has a strong dependence on TS feature definition which varies from different application scenarios. Considering the challenges in Data Streams: Feature Evolution, Concept Evolution and Concept Drift, we gave several contexts within two stream definitions: Time Series Stream and Streaming Time Series. We showed as well our preliminary work on adaptive feature exploration over TS Stream considering Concept Drift.

Future work is planned on both TS Stream and Streaming TS context. Firstly, system efficiency and caching mechanism are two mains aspects to optimise for TS Stream. As for the context of Streaming Time Series, complex human activity recognition and active learning on weak-labelled data are on our to-do list, where the interaction between system and human is the research focus.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. 2010. *MOA: Massive Online Analysis.* Technical Report. 1601–1604 pages.
[2] Mustafa Gokce Baydogan, George Runger, and Eamonn B Keogh Mustafa Gokce Baydogan. 2016. Time series representation and similarity based on local autopatterns. *Data Mining and Knowledge Discovery* 30 (2016), 476–509.
[3] Ahsanul Haque, Latifur Khan, and Michael Baron. 2016. SAND: Semi-Supervised Adaptive Novel Class Detection and Classification over Data Stream. *AAAI Conference on Artificial Intelligence* (2016), 1652 – 1658.
[4] Jason Lines, Anthony Bagnall, J Lines, · A Bagnall, and A Bagnall. 2015. Time series classification with ensembles of elastic distance measures. *Data Min Knowl Disc* 29 (2015), 565–592.
[5] Jason Lines, Jon Hills, and Anthony Bagnall. 2015. The Collective of Transformation-Based Ensembles for Time-Series Classification. *IEEE Transactions on Knowledge and Data Engineering* 27, 9 (2015), 2522–2535.
[6] J. Lines, S. Taylor, and A. Bagnall. 2016. HIVE-COTE: The Hierarchical Vote Collective of Transformation-Based Ensembles for Time Series Classification. In *2016 IEEE 16th International Conference on Data Mining (ICDM).* 1041–1046.
[7] Li Liu, Yuxin Peng, Shu Wang, Ming Liu, and Zigang Huang. 2016. Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors. *Information Sciences* 340-341, January (2016), 41–57.
[8] Alice Marascu, Suleiman A. Khan, and Themis Palpanas. 2012. *Scalable similarity matching in streaming time series.* Technical Report PART 2. 218–230 pages.
[9] George Runger Mustafa Gokce Baydogan and Eugene Tuv. 2017. A Bag-of-Features Framework to Classify Time Series. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 39, 10 (2017), 2104–2111.
[10] P. Senin and S. Malinchik. 2013. SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model. In *2013 IEEE 13th International Conference on Data Mining.* 1175–1180.
[11] Lexiang Ye and Eamonn Keogh. 2009. Time series shapelets: A New Primitive for Data Mining. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09* (2009), 947.
[12] Jingwei Zuo, Karine Zeitouni, and Yehia Taher. 2019. Exploring Interpretable Features for Large Time Series with SE4TeC. In *Proc. EDBT 2019.* 606–609.