M2 RESEARCH INTERNSHIP

STF-Online: Adding Spatio-temporal Forecasting to STREAMER Framework

Keywords: *Artificial Intelligence, Spatio-temporal data, Online Forecasting, Data Streams, Concept Drift, Meta-learning*

With the development of smart cities, sensors are becoming ubiquitous, and allow monitoring the *city pulse*. The geo-located sensors and their generated sequential data (i.e., time series) represent a complex spatio-temporal structure. Applications ranging from climate or air pollution studies to transportation planning require advanced analysis of these data, using machine learning. **Spatio-Temporal Forecasting (STF)** is one typical example of such learning tasks.

However, a conventional STF model trained on static dataset does not hold in the dynamic streaming context because of **Concept Drift**, i.e., the data distribution changes with time, or more specifically, the non-stationarity in the *streaming spatio-temporal data*. Therefore, the STF model should adapt to the dynamic context where the streaming spatio-temporal data is generated in real-time. This task is challenging due to the complex spatio-temporal dependencies, namely, the temporal dynamics in long-(short-) term [1], the spatial correlation between nearby sensors [2], and the correlation between the sensors (of different parameters) at the same site. The challenge also comes from the lack of ground truth and the difficulty of interpretation of the possible root causes of a drift.

Spatio-temporal forecasting has been studied for decades. Beside Auto-Regressive Integrated Moving Average (ARIMA) and Kalman filters models for time series forecasting, deep learning models for STF have been developed recently in [2] [3], proving the strong performance of the Graph Convolutional Networks (GCNs) on modeling the complex spatio-temporal structure, but without considering the concept drift (i.e., non-stationarity), nor the streaming spatio-temporal context. While the non-stationarity problem in STF has been addressed in [4], the model updates were costly and unsuitable for the streaming context. Most works have focused on drift detection, but without performing model selection as they are typically limited to a single model [5]. Meta-learning has been proposed as a way to perform model selection automatically. Its performance has been recently demonstrated in the context of time-series forecasting [6]. However, the complex structure of streaming spatio-temporal data requires a thorough re-design of the forecasting model on **drift detection**, **drift interpretation** and **model selection**.

Besides, we have developed in our previous work **STREAMER** [7], a data stream processing framework for integrating and testing machine learning algorithms in realistic streaming operational contexts. It allows data scientists abstracting from the implementation of the stream to only focus on their use cases (involving algorithms, data pre-processing and evaluation functions) by adding them or simply reusing the existing ones. STREAMER framework can be deployed in any operating system, accepts the integration of algorithms programmed in a wide variety of programming languages, provides two graphical interfaces, and is free and accessible for everybody. Its code is published in open source under GNU3 license and counts with its official website: https://streamer-framework.github.io

Therefore, based on our previous work on data streams [5] [7] [8], in the framework of the DATAIA StreamOps project, the **ultimate objective of this internship is two-fold**:

- First, to propose an online STF model considering the drift detection, drift interpretation and model selection in the spatio-temporal context.
- Then, to integrate the online STF model into STREAMER [9], the open-source data stream processing framework, and test it over public datasets[1].

Profile and skills required

The applicant should be currently a 2nd year Master student or in the last year of engineering school in Computer Science. She/he should have:

- Strong background in machine/deep learning
- Strong object and system programming skills
- Good English oral communication, technical reading and writing skills
- Collaboration skills in general
- Proficiency in French is desirable but not mandatory

Hosting team and laboratory. The internship will take place at ADAM team, in the laboratory *Données et Algorithmes pour une Ville Intelligente et Durable* (DAVID) - UVSQ, Université Paris-Saclay, Versailles. The intern will have a close collaboration with CEA-LIST for the model implementation and integration into STREAMER framework. He will be advised by the research team of StreamOps:

- Karine Zeitouni, DAVID/UVSQ, Université Paris-Saclay, Professor (main-advisor)
- Sandra Garcia Rodriguez, CEA-LIST, Research Engineer
- Jingwei Zuo, DAVID/UVSQ & Mohammad Alshaer, CEA-LIST (resp. PhD candidate and Postdoc)

Duration and benefits. Funded by DATAIA institute, the internship will start from March/April 2021 for 5 to 6 months paid around 600 euros per month.

Application. The applicant must send a resume, motivation letter, academic transcripts, recommendation letters (if any) to karine.zeitouni@uvsq.fr and sandra.garciarodriguez@cea.fr

References

1. Lai G., Chang W., Yang Y., and Liu H., Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks, SIGIR'18
2. Li Y., Yu R., Shahabi C., and Liu Y., Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting, ICLR'18
3. Song C., Lin Y., Guo S., and Wan Y., Spatial-Temporal Synchronous Graph Convolutional Networks: A New Framework for Spatial-Temporal Network Data Forecasting, AAAI'20
4. Matsubara Y. and Sakurai Y., Regime Shifts in Streams: Real-time Forecasting of Co-evolving Time Sequences, KDD'16
5. Zuo J., Zeitouni K, and Taher Y., Incremental and Adaptive Feature Exploration over Time Series Stream. IEEE Big Data 2019
6. Candela R., Michiardi P., Filippone M., and Zuluaga M. A., Model Monitoring and Dynamic Model Selection in Travel Time-series Forecasting, ECML-PKDD'20
7. Sandra GR., Alshaer M., and Gouy-Pailler C., STREAMER: a Powerful and Open-Source Framework for Continuous Learning in Data Streams, demo, CIKM'20
8. Alshaer M., Sandra GR. and Gouy-Pailler C., Detecting Anomalies from Streaming Time Series Using Matrix Profile and Shapelets Learning. ICTAI'20

[1] e.g., the dataset published in Array Of Things project: https://arrayofthings.github.io/node-locations.html