PHD PROPOAL

Université Paris-Saclay Doctoral Programme in Artificial Intelligence

DISTRIBUTED DATA STREAM LEARNING IN A COLLABORATIVE ENVIRONMENT

Keywords: Machine Learning, Data Stream, Multi-agent Systems, Distributed Data Processing, Data Mining

With the advance of technology, the last decade saw the rise of applications generating a large quantity of data in a continuous way [1]. Such data may come from different sources and domains as social media exchanges, online transactions, sensors, GPS signals, mobile phones and many other connected objects. This comes with a proliferation of solutions for the collection and analysis of such *data streams* [12] for various purposes. Learning from this unbounded and infinite arrival of data is crucial. Nevertheless, since storage devices are limited, classic big data approaches are not suitable for a full knowledge extraction since they require, as batch-processing oriented methods, the entire historical database. Thus, the challenge about how to overcome this limitation emerges. There is a need for powerful analysis tools and algorithms that are able to process and learn from continuous data streams [3] without storing them. Formalized by Muthukrishnan [2], **data streams** are defined as infinite streams of data integrated from both live and historical sources. In such scenarios, data stream processing algorithms should satisfy requirements as bounded storage, single pass (data is going to be processed just once), real time and concept drift [13-15].

Nowadays, data streams are present in more and more applications and domains where dynamism and speed truly matter. In practice those streams represent dynamic data flows, coming from different sources, where their content evolves in time. Research has been done in the subject and many techniques for stream mining have emerged [3]. These algorithms usually sample the data stream in a certain way and deal with them incrementally or online. Despite the results provided by these new techniques, the flow of data is still underutilized, which potentially leads to a loss of useful information on the one hand, and to forget what has been previously discovered on the other. Moreover, the complexity of current digital applications, and those of the near future, is constantly increasing due to a combination of aspects such as the large number of sources, the non-linearity of certain processes, the distribution of knowledge and control, the time response, the strong dynamics of its environment or the unpredictability of interactions among others.

The aforementioned complexity opens new research challenges about the generation and processing (learning) of these streams, especially in distributed, heterogeneous and collaborative environments. Existing ones lack, in general, the means for collaborating, negotiating, sharing, or validating data streams on such kind of heterogeneous environments. Multi-Agent Systems (MAS) have been demonstrated to be an appropriate technology for dealing with those issues. Their principles enable some of these features but however there is still work to do in order to comply them with the characteristics of data streams.

But distributed data stream solutions also require distributed data stream learning algorithms [11]. Data stream learning has been attracting a lot of research efforts recently, however, this problem has not received enough consideration when the data streams are generated in a distributed fashion, whereas such scenario is very common in real life applications. On the other hand, distributing the workload of the algorithm among different nodes helps improving its scalability.

The goal of this PhD is then to propose solutions to the aforementioned challenges, and contribute in the field of data stream machine learning. More precisely we pursue the following objectives:

i. Propose agent-based solutions for managing non-synchronized and distributed data streams. We want to explore how multi-agent systems could handle data processing in a continuous manner (as plugging stream reasoners into the processing core of an agent). We want to use their goal-oriented nature to help improving the processing efficiency, response time boundaries; and to manage continuous query requirements. Their intrinsically reactive nature could generate reactive responses to interactions exchanged by (groups of) agents. Also their capability to capture knowledge and incorporate it to their beliefs would allow them to create a context of the current situation and act accordingly.
ii. Propose distributed learning algorithms applied to streaming context. Unlike traditional data mining, data stream mining is a continuous learning process while coping with time and memory limitations. Scalability and concept drift adaptation are two key issues of data stream mining. To address concept drifts, data stream algorithms need novel strategies to efficiently detect them at varying time windows. In order to achieve scalable performance, a data stream algorithm must minimize the number of passes over the data while fitting the data synopsis into the main memory. Another way for improving scalability in algorithms is to share their workload among different parallel nodes instead of just relying in a single one, in other words, making them work in a distributed manner.
iii. Generate new metrics to evaluate the features mentioned before.

The implementation of those solutions and their metrics will rely and extend STREAMER. Developed in LI3A laboratory (CEA List) premises, STREAMER is a cutting-edge data stream processing (Complex Event Processing) framework devoted to analyzing sequential data for electrical or industrial systems.

Profile and skills required

The applicant should hold a Master diploma in Computer science, or equivalent. She/he should have: -
Strong object and system programming, and database skills
- Good background in machine learning
- Good English oral communication, technical reading and writing skills
- Proficiency in French is desirable but not mandatory

Details on the thesis supervision

This thesis will be co-supervised at 70% in CEA and 30% in UVSQ. The candidate will be hosted in CEA LIST premises in Saclay 4 days per week and in UVSQ – DAVID Lab (Versailles) 1 day a week.
Co-advisers:

Sandra GARCIA RODRIGUEZ (CEA), Research Engineer sandra.garciarodriguez@cea.fr
Karine ZEITOUNI (UVSQ), Professor (Thesis Director) karine.zeitouni@uvsq.fr

Application Procedure

This proposal is in the framework of the Université Paris-Saclay Doctoral Programme in Artificial Intelligence UDOPIA, co-funded by the National Research Agency (ANR). Interested candidates should provide detailed CV, a motivation letter, transcripts of the last academic year, and can attach recommendation letters. Submission website:
http://www.adum.fr/as/ed/voirproposition.pl?site=PSaclay&matricule_prop=31677&langue=en

Application deadline : May 19, 2020

It's also required that they send their application the advisors before the deadline (see above).

REFERENCES

[1] Kolajo, T., Daramola, O., Adebiyi, A.: Big data stream analysis: a systematic literature review. Journal of Big Data6(1), 47 (2019).

[2] Muthukrishnan, S. (2005). Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, *1*(2), 117-236.
[3] Palpanas, Themis. "Data series management: The road to big sequence analytics." *ACM SIGMOD Record* 44.2 (2015): 47-52.
[4] Tommasini, Riccardo, Davide Calvaresi, and Jean-Paul Calbimonte. "Stream Reasoning Agents: Blue Sky Ideas Track." *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2019.
[5] Xie, Jing, and Chen-Ching Liu. "Multi-agent systems and their applications." *Journal of International Council on Electrical Engineering* 7.1 (2017): 188-197.
[6] Krempl, Georg, et al. "Open challenges for data stream mining research." *ACM SIGKDD explorations newsletter* 16.1 (2014): 1-10.
[7] Twardowski, Bartlomiej, and Dominik Ryzko. "Multi-agent architecture for real-time big data processing." *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Vol. 3. IEEE, 2014.
[8] Belghache, Elhadi, Jean-Pierre Georgé, and Marie-Pierre Gleizes. "Towards an adaptive multi-agent system for dynamic big data analytics." *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. IEEE, 2016.
[9] Nguyen, Hai-Long, Yew-Kwong Woon, and Wee-Keong Ng. "A survey on data stream clustering and classification." *Knowledge and information systems* 45.3 (2015): 535-569.
[10] Babcock, Brian, et al. "Models and issues in data stream systems." *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2002.
[11] Peteiro-Barral, Diego, and Bertha Guijarro-Berdiñas. "A survey of methods for distributed machine learning." *Progress in Artificial Intelligence* 2.1 (2013): 1-11
[12] Isah, Haruna; Abughofa, T.M.S.A.D.Z.F., Khan, S.: A survey of distributed datastream processing frameworks. IEEE Access7, 154300–154316 (2019)
[13] Zuo, J., Zeitouni, K., Taher, Y.: Incremental and Adaptive Feature Exploration over Time Series Stream. IEEE International Conference on Big Data in 2019.
[14] Zuo, J., Zeitouni, K., Taher, Y.: ISETS: Incremental Shapelet Extraction from Streaming Time Series. ECML/PKDD 2019 [Demo].
[15] Zuo, J., Zeitouni, K., Taher, Y.: Incremental and Adaptive Feature Exploration over Time Series Stream. AALTD, joint workshop to ECML/PKDD 2019.