TensAIR: Online Learning from Data Streams via Asynchronous Iterative Routing

Mauro Dalle Lucca Tosi, Martin Theobald

Department Of Computer Science, University of Luxembourg, Luxembourg
{mauro.dalleluccatosi,martin.theobald}@uni.lu

Online Learning (OL) is a branch of *Machine Learning* (ML) which studies solutions to time-sensitive problems that demand *real-time answers* based on fractions of data received from *data streams*. Despite recent advances in the computational infrastructures, training complex OL models, such as those based on *Deep Neural Networks* (DNNs), in real-time remains a major challenge. Current stream-processing platforms, such as Apache Kafka and Flink, already come with basic extensions for training DNNs from data streams; however, these extensions were not originally designed to train DNNs in real-time, and they still suffer from various performance- and scalability-related issues when doing so.

TensAIR Architecture: In this talk, we present an overview of TensAIR [1], one of the first OL systems which has been designed from scratch for training DNNs in real-time. TensAIR achieves remarkable performance and scalability gains when training DNN models (either freshly initialized or pre-trained) via a form of *decentralized and asynchronous stochastic gradient descent* (DASGD) [2]. We demonstrate the versatility of TensAIR over both sparse (word embeddings) and dense (image classification) use-cases, for which TensAIR is able to achieve from 6 to 116 times higher throughput rates than state-of-the-art systems.

OPTWIN Drift Detector: We additionally present OPTWIN [3], our "OPTimal WINdow" *concept-drift detector* suited for classification and regression problems. OPTWIN uses a sliding window of events over an incoming data stream to track the prediction errors of an OL algorithm. OPTWIN is able to split such a sliding window of error rates into two *provably optimal sub-windows*, such that the split occurs at the earliest event at which a statistically significant difference according to either the t- or the f-test occurs. Moreover, by combining both test statistics into a single objective function, it is able to determine this optimal split in constant ($\mathcal{O}(1)$) time, regardless of the window size. We assessed OPTWIN on the common MOA framework over no less than 5 concept-drift baselines and 12 (both synthetic and real-world) datasets. OPTWIN surpasses the F1-score of the baselines by maintaining a lower detection delay and saving up to 21% of time spent on retraining the models.

References

[1] Tosi, M. D. L., Venugopal, V. E., & Theobald, M. (2022). TensAIR: Online Learning from Data Streams via Asynchronous Iterative Routing. *arXiv preprint arXiv:2211.10280*.

[2] Tosi, M. D. L., & Theobald, M. (2023). Convergence Analysis of Decentralized ASGD. *arXiv preprint arXiv:2309.03754*.

[3] Tosi, M. D. L., & Theobald, M. (2023). OPTWIN: Drift identification with optimal sub-windows. *arXiv preprint arXiv:2305.11942*.