Ph.D. in Information Technology
Thesis Defense

May 24th, 2023
at 15:00
Aula PT1

Alessio BERNARDO – XXXV Cycle

ON LEARNING FROM MASSIVE, EVOLVING, AND IMBALANCED DATA STREAMS

Supervisor: Prof. **Emanuele Della Valle**

Abstract:

Data are everywhere. From new emerging topics in social networks to pressure and vibration levels of industrial machinery, from traffic congestion to drilling in an oil ring, data streams are now more than ever an invaluable resource for companies, if only they could correctly analyze them.

Indeed, around 68% of the data generated are not used due to inadequate infrastructures and techniques. Most of the traditional Machine Learning techniques used by companies cannot deal with a continuous coming flow of data due to how they manage their learning phase. More specifically, they restart anew the training process every time new data are available (stateless retraining), resulting inappropriate in a world where unboundedness, velocity, volatility, and non-stationarity are the new normal.

Recently, a new generation of Machine Learning models, called Streaming Machine Learning, was introduced to cope with the new data streams requirements, i.e., processing data on the fly (one sample at a time or in mini-batches), incrementally and continuously updating the models (stateful retraining), quickly adapting to any non-stationarity change in the stream (concept drift), and discarding the samples after updating the models to manage the time and memory consumption.

Although Streaming Machine Learning can manage data streams, some challenges remain unsolved. In a binary classification scenario, one of them is the combination of multiple concepts drift occurrences over time with class imbalance, from which the following unsolved research question arose: *In case of imbalanced data streams and concept drifts, in a binary classification task, is it possible to conceive novel rebalancing meta-strategies that help in outperforming the state-of-the-art?*

This PhD thesis deeply investigates the class imbalance problem in the streaming scenario proposing new meta-strategies to be combined with any SML model unable to rebalance streams in the presence of concept drifts. The investigation narrows the principal challenge into subsequently more specific ones, aiming at better investigating and highlighting the fundamental aspects that a meta-strategy must fulfil.

Finally, this thesis, starting from the narrowest challenges, wraps up the results achieved by all the meta-strategies proposed, presenting benefits and limitations, and uses them to discuss the more general challenges up to the principal research question.

PhD Committee

Prof. **Giacomo Boracchi,** Politecnico di Milano
Prof.**Davide Bacciu**, Università di Pisa
Prof. **Danh Le Phuoc**, Technical University of Berlin