ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

ΔΗΜΟΚΡΙΤΟΣ
DEMOKRITOS

Προτεινόμενα Θέματα Διπλωματικής Εργασίας

Από

Νίκο Κατζούρη

A. Learning Finite State Machines for Complex Event Detection & Forecasting

Συνοπτική περιγραφή:

Finite State Machines (FSMs), also called *automata*, are expressive graph-based representations of sequential patterns. An FSM may be used to describe processes, protocols, plans, policies in reinforcement learning[1] and in principle, any kind of time-spanning pattern.

Techniques for *complex event detection and forecasting* may use such FSM-based patterns to monitor, across datastreams, interesting situations as they evolve in time, in order to efficiently identify, or even forecast[2] critical occurrences of events. However, such patterns are not always known beforehand and even if they are, they often need to be updated in order to reflect change in the underlying data distributions, or drift in the monitored concepts (i.e. the target complex events that are to be detected or forecast). Therefore, it is often necessary to bring machine learning techniques to the rescue, which may automatically discover such patterns from data.

If the complex event patterns are represented by FSMs, as it is frequently the case, such machine learning techniques are collectively known as “automata induction” techniques and they have a long history[3]. However, given that the problem of FSM learning is computationally hard, most such techniques use heuristically-driven algorithms that tend to learn over-complicated, frequently overfitted models that generalize poorly to unseen data and are hard to inspect and interpret. One means

[1] Hasanbeig, M. et al., DeepSynth: Automata Synthesis for Automatic Task Segmentation in Deep Reinforcement Learning, AAAI, 2021

[2] Alevizos E. et al., Event Forecasting with Pattern Markov Chains, DEBS 2017

[3] De la Higuera, Colin. Grammatical inference: learning automata and grammars. Cambridge University Press, 2010.

towards simplifying such models, in addition to developing better learning algorithms, is to target so-called *symbolic automata*. Transitions in such automata are labelled with symbolic expressions, rather than mere symbols, and that can significantly increase the expressive power of the model while drastically reducing its complexity.

The topic of this thesis is to study FSM learning techniques addressing such issues, i.e. improving efficiency and scalability of learning (near-) optimal FSMs via incremental and online learning techniques, and improving the expressive power of the learnt models via symbolic automata learning.

The starting point will be existing algorithms and software, which will be both extended accordingly to achieve (some) of the aforementioned goals. The project requires a good handle of Python and machine learning basics, while some familiarity with Answer Set Programming will be helpful (although not a prerequisite).

Experiments will be carried-out on multivariate time series-like data from real-life domains such as cancer cell evolution simulations, human activity recognition and maritime surveillance applications.

B. **Neural-Symbolic Answer Set Programming**

Συνοπτική περιγραφή:

Machine learning and machine reasoning have been largely addressed separately and in isolation by different communities in Artificial Intelligence. *Learning* traditionally refers to data-driven, subsymbolic techniques for generating predictive models and it is frequently related to low-level (e.g. perception-level) tasks, especially within deep learning. *Reasoning,* on the other hand, refers mostly to symbolic, frequently logic-based techniques for deriving new knowledge from data and existing knowledge, and it is typically associated with higher-level inference tasks. Artificial Intelligence really needs both learning and reasoning in order to bring to life systems that combine the best of two worlds, i.e. systems capable of perceiving their environment and making sense of data, while also reasoning with what has been learnt and consulting existing knowledge about the world.

Neural-symbolic computation[4] in particular, seeks to integrate neural-based (deep) learning with logic-based reasoning and it has the potential of addressing many of the shortcomings of contemporary AI approaches, including the black-box nature and the brittleness of deep learning, and the difficulty to adapt knowledge representation models in the light of new data.

The topic of this thesis is related to methods that integrate neural and symbolic computation, based on Answer Set Programming (ASP)[5]. The latter is a declarative problem solving methodology oriented towards solving combinatorial optimization

[4] Garcez A. et al, Neural-symbolic computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. FLAP, 2019

[5] Lifschitz, V. Answer set programming. Springer, 2019.

problems. ASP offers mature and sophisticated tools (answer set solvers) capable of dealing with various AI tasks, such as complex reasoning under uncertainty and symbolic learning. The starting point for this thesis will be existing approaches[6] and software[7], based on a combination of Keras and the Clingo ASP solver, towards combining deep learning with symbolic reasoning. Part of the thesis involves understanding the starting-point approach, working with the existing software, trying it in different settings and with different datasets and possibly extending it to address properly identified challenges. The project requires a good handle of Python and deep learning basics. Some familiarity with Keras and ASP will be helpful (although not a prerequisite).

C. Mining Highly-Informative, Rare Patterns from Datastreams Under Scarce Supervision

Συνοπτική περιγραφή:

Dealing with the abundance of dynamic data flows in contemporary applications calls for scalable learning algorithms capable of making sense of such data by extracting patterns. Typically, supervised learning techniques are used by such algorithms, where portions of a datastream that refer to such interesting patterns are labelled beforehand by human annotators. However, this process is labor-intensive, error-prone and expensive. Moreover, interesting patterns, especially rare, but highly informative ones are not always known. It is therefore necessary to develop machine learning techniques capable of discovering such rare and interesting patterns using extremely scarce (or even none at all) supervision in terms of ground truth (labels).

It is also necessary for the extracted patterns to be as transparent as possible, so that human experts can associate them with actual situations and informative domain features by simply inspecting them.

As examples of such rare but informative situations consider motifs describing sea piracy incidents, a car crash in a highway, a DDOS attack to a computer network, a sudden burst (or collapse) in the price of a cryptocurrency, or a sequence of odd/unexpected measurements from the machinery that monitors a patient in a hospital.

The topic of this thesis is to study machine learning techniques for extraction of highly informative, often rare, temporal and/or sequential patterns from datastreams under limited supervision.

The starting point will be existing learning techniques for pattern mining from data[8]

[6] Yang et al, Neurasp: Embracing Neural Networks Into Answer Set Programming, IJCAI 2020.

[7] https://github.com/zhunyoung/NeurASP

[8] Yang Y, Chen L, Fan C. (2021) ELOF: fast and memory-efficient anomaly detection algorithm in data streams. *Soft Comput.*, 25(6):4283-4294.

[9] [10] [11], which will be extended accordingly to achieve (some) of the aforementioned goals.

The project requires a good handle of Python and basic knowledge of machine learning. Experiments will be carried-out on multivariate time series-like data from real-life domains such as cancer cell evolution simulations, human activity recognition and maritime surveillance applications.

[9] Fouché E, Kalinke F, Böhm K. (2021) Efficient Subspace Search in Data Streams. *Information Systems*.

[10] Zubaroğlu A, Atalay V. (2020) Data stream clustering: a review. *Artif Intell Rev.*, 54(2), pp. 1201-1236.

[11] Liu FT, Ting KM, Zhou Z-H. (2008) Isolation Forest. In: *Proceedings of the 8th IEEE International Conference on Data Mining*. IEEE Computer Society, pp. 413-422.