**Éligible à une allocation de type :**

**Titre du sujet :**

**Mention de thèse :**

**HDR Directeur de thèse inscrit à l'ED STIC :**

---

**Co-encadrant de thèse éventuel :**

**Nom :**

**Prénom :**

**Email :**

**Téléphone :**

**Email de contact pour ce sujet :**

**Laboratoire d'accueil :**

---

**Description du sujet :**

Recent advances in Machine Learning (ML) have brought new solutions for the problems of prediction, decision, and identification. ML is impacting almost all domains of science or industry but determining the right ML workflow for a given problem remains a key question. To allow not only experts in the field to benefit from ML potential, last years have seen an increasing effort from the big data companies (Amazon AWS, Microsoft Azure, Google AutoML...) to provide any user with simple platforms for designing their own ML workflow. However, none of these solutions consider the design of ML workflow as a generic process intending to capture common processing patterns between workflows (even through workflows targeting different application contexts). These platforms either propose a set of dedicated solutions for given classes of problem (i.e. AutoML Vision, AutoML natural language , AutoML Translation...) or propose a recipe to build your own ML workflow from scratch (i.e. MS Azure Machine Learning studio, RapidMiner).
Is it then possible to envision the meta-learning process of designing ML workflow as a systematic approach analyzing past experiences to identify, explain and predict the right choices?

This PhD thesis will address this issue by correlating research on software architectures (including product lines) and meta-learning, to bring ML workflow design to the next level by producing explanation on algorithm choices and by cutting portfolio exploration complexity identifying common patterns between workflows.

Context: To determine the right ML workflow for a given problem, numerous parameters have to be taken in account: the kind of data, expected predictions (error, accuracy, time, memory space), the choice of the algorithms and their judicious composition [14,11]. Faced with the complexity of choosing the "right" assembly, meta-learning offers an attractive solution, learning from the problems of the past. The algorithm selection problem is one of its applications [5]: given a dataset, identify which learning algorithm (and which hyperparameter setting) performs best on it. Algorithm Portfolio generalizes the problem and automates the construction of selection models [8]. The immediate goal is the same: to predict the results of the algorithms on a given problem without executing them. Even if, in the portfolio, some selection models are built by meta-learning [1], the purpose is different. The portfolio is based on the systematic acquisition of knowledge about the algorithms it contains. The research then focuses on the quality and the return of knowledge, the acquisition process itself and the construction of selection models over time. See link below for more information.

Objectives: The construction of a portfolio requires covering a space of experiments broad enough to "cover" all the problems that may be submitted to it. However, (1) the space of problems and solutions presents a very great* “diversity” [16] even within a single class of problem like classification [9]. (2) The resources required for ML experiments are massive (time, memory, energy)**. (3) As the ML domain is particularly productive, the portfolio must be able to evolve to integrate new algorithms. (4) To cope with the mass of data, the transformation of experimental results into knowledge requires the implementation of automatic analysis procedures.

The PhD work will be organized to provide contributions in the following directions:

1- A representation of experiments in the form of graphs [10] and exploitation of these structures by adapted learning algorithms[3,15]; In particular, this approach should be exploited to reduce the search space and explain the choices made[4];

2- An automation of the analysis of the results of experiments to identify properties and patterns of workflows such as similarities, cliques/stables that correspond to algorithms that statistically must always or never be used together;

3- A systematic exploitation of this structure to reduce the number of executions, to drive the workflow compositions, to manage the feedback loop, and to justify choices.

Bibliographie

URL: https://mireilleblayfornarino.i3s.unice.fr/doku.php?id=students:phd_2019

English version: