Ecole Doctorale STIC

Sciences & Technologies de l'Information et de la Communication

MINES ParisTech

Université Nice SOPHIA ANTIPOLIS

ED STIC - Proposition de Sujets de Thèse

pour la campagne d'Allocation de thèses 2017

Axe Sophi@Stic : aucun|

Titre du sujet : Machine Learning Workflow System

Mention de thèse : Informatique

HDR Directeur de thèse inscrit à l'ED STIC : Mireille Blay-fornarino

Co-encadrant de thèse éventuel :

Nom :

Prénom :

Email :

Téléphone :

Email de contact pour ce sujet : Mireille.Blay@unice.fr

Laboratoire d'accueil : I3S

Description du sujet :

Context

For many years, Machine Learning research has been focusing on designing new algorithms for solving similar kinds of problem instances (Kotthoff, 2016). However,

Researchers have long ago recognized that a single algorithm will not give the best performance across all problem instances, e.g. the No-Free-Lunch-Theorem (Wolpert,

1996) states that the best classifier will not be the same on every dataset. Consequently, the “winner-take-all” approach should not lead to neglect some algorithms that,

while uncompetitive on average, may offer excellent performances on particular problem instances. In 1976, Rice characterized this as the "algorithm selection problem"

(Rice, 1976).

To support automatic selection of algorithms, Portfolio approaches aim at performing per-instance algorithm selection (Leyton et al., 2003). When portfolio refers to more complex products than algorithms (i.e. not only a set of software components but the composition of a set of consistent software components), Software Product Line (SPL) is a successful approach to increase the product portfolio with up to an order of magnitude and provide consistent user experience across product portfolio (Bosch J, 2009).

Software Product Line engineering is concerned with systematically reusing development assets in an application domain (Clements et al., 2001)(Pohl et al., 2005).

A Machine Learning (ML) Workflow can be defined as a tuple (h,p,c) where h represents hyper-parameter tuning strategy, p represents a set of preprocessing techniques applied on the dataset, and c is a ML algorithm used to learn a model from the processed data and to predict then over new data. The construction of a Machine Learning Workflow depends upon two main aspects: (1)The structural characteristics (size, quality, and nature) of the collected data, (2) How the results will be used.

This task is highly complex because of the increasing number of available algorithms, the difficulty in choosing the correct preprocessing techniques together with the right algorithms as well as the correct tuning of their parameters. To decide which algorithm to choose, data scientists often consider families of algorithms in which they are experts, and can leave aside algorithms that are more “exotic” to them, but could perform better for the problem they are trying to solve.

ROCKFlows is a project aiming at helping users to create their own Machine Learning Workflows by simply describing their dataset and objectives. The approach is thus positioned differently from big companies approaches or from the platforms that help select the workflows components (See detailed description).

Objectives

The main objective of this thesis is to explore the alliance between a portfolio and a SPL to automatically propose ML workflows according to end-user problems. So the SPL is the link between the portfolio and the end-user. It manages the identification of the end-user problem. It proposes solutions among which end-user chooses according to her own criteria. It generates the corresponding codes and, it could launch the experiment. It must be able to collect the results of the experiments to get feedbacks and eventually to enrich the platform.

The thesis must address the following challenges: Relevance and quality of predictions and Scalability to manage the huge mass of ML workflows.

To meet these challenges, attention should be paid to the following aspects:

- Handling Variabilities: Variability of compositions (e.g. identifying dominated workflows, managing requirements between WF components); Variability of performance metrics (e.g. dependencies among metrics); (See detailed description).
- Architecture of portfolio to automatically manage (1) experiment running, (2) collect of experiment results, (3) analyze of results, (4) evolution of algorithm base. It must support the management of execution errors, incremental analyzes, identifying context of

experiments.
Handling Scalability of Portfolio: Selecting discriminating data sets; Detecting “deprecated” algorithms and WF from experiments and literature revues; Dealing with information from scientific literature without deteriorating portfolio computed knowledge.
Ensuring global consistency of Portfolio and Software Product Line. Such a system is enriched by additions to the portfolio and experiment feedbacks. As "knowledge" evolves (e.g., new data types, new metrics), the entire system needs to be updated. It is therefore to find abstractions not only to manage these changes but also to optimize them (Bischl et al. 2016).

We have a two-year experience on this subject which has enabled us to (I) eliminate some approaches (e.g. modeling knowledge as a system of constraints because it generates on our current basis more than 6 billion constraints), (ii) lay the foundations for a platform for collecting experiences and presenting to the user (Camillieri et al., 2016) (see http:// http://rockflows.i3s.unice.fr/), (iii) study the ML workflows to predict workflows (Master internships Luca Parisi, Miguel Fabian Romero Rondon and Melissa Sanabria Rosas), (iv) address platform evolution introducing deep learning workflows (see Melissa’s Report).

The thesis must investigate the research around the selection of algorithms, considering the automatic composition of workflows and supporting dynamic evolutions. It is therefore a thesis in software engineering research but to address one of the current most central problems in machine learning.

Bibliographie
See https://mireilleblayfornarino.i3s.unice.fr/students:phd_mlws

URL : https://mireilleblayfornarino.i3s.unice.fr/students:phd_mlws

English version: