WOLF: Machine Learning Workflow Management Framework

Authors Name/s per 1st Affiliation (Author)
line 1 (of Affiliation): dept. name of organization
line 2: name of organization, acronyms acceptable
line 3: City, Country
line 4: Email: name@xyz.com

Authors Name/s per 2nd Affiliation (Author)
line 1 (of Affiliation): dept. name of organization
line 2: name of organization, acronyms acceptable
line 3: City, Country
line 4: Email: name@xyz.com

***Abstract*—The abstract goes here. DO NOT USE SPECIAL CHARACTERS, SYMBOLS, OR MATH IN YOUR TITLE OR ABSTRACT.**

***Keywords*-Machine Learning Workflow; Automatic Machine Learning;**

I. Introduction

II. Related Work

In recent years, various platforms and techniques have been developed for easing cumbersome tasks in a machine learning (ML) workflow, such as data preprocessing, feature extraction and selection, and model selection. They can be categorized into two categories, ML workflow management and ML workflow discovery.

A. ML Workflow Manaagement

Platforms and techniques in this category provide tools for creating, modifying, executing, or sharing ML workflows. The FBLearner Flow system from Facebook was designed to be capable of easily reusing algorithms, scaling to run thousands of simultaneous custom experiments, and managing experiments with ease. KNIME enables easy visual assembly and interactive execution of a data pipeline through customizable and extentionable nodes. The Portable Format for Analytics is an emerging standard for statistical models and data transformation engines. PFA combines portability across systems with algorithmic flexibility: models, preprocessing, and post-processing are all represented as functions that can be arbitrarily composed, chained, or built into more complex workflows. Kepler provides an graphical interface for creating a "scientific workflow"an executable representation of the steps required to generate results.

B. ML Workflow Discovery

The techniques in the second category is more related to WOLF in the sense that they aim to discover an optimal ML workflow for a ML task. TPOT is a Python automated machine learning tool that optimizes machine learning workflows using genetic programming. The scalability of TPOT may be problematic since its process of finding optimal workflow were not designed for distributed computation. To addressing the issue of scalability in machine learning, Tim et al. proposed MLBase framework consisting of three components, ML Optimizer, MLI, and MLLib. Through ML Optimizer, an optimal learning plan can be selected for a ML task, such as classification, specified using a declarative language. Note that ML Optimizer is still under development and MLLib is specifically designed for Spark. It requires non-negligible efforts to implement an Spark compatible algorithm to achieve distributed computation. DataRobot is a proprietary data science system with software that covers the tasks that a data scientist typically performs. It is designed to automate the task of data cleaning, visualization, model construction, model evaluation, and making predictions.