Bachelor- / Master-Thesis: Integration eines AutoML-Frameworks in ein Semantisches Data-Lake-System

Motivation

Data-Lake-Systeme werden seit einigen Jahren als Repositories vorgeschlagen, in denen heterogene Daten gespeichert und zusammengeführt werden können [1]. Für die Entwicklung solcher Systeme müssen verschiedene Technologien aus den Bereichen Big Data, Datenbanken und maschinelles Lernen kombiniert werden. In den letzten Jahren wurden in mehreren studentischen Projekten verschiedene Elemente eines Data-Lake-Systems namens *SEDAR* entwickelt, die nun in die Produktionsumgebung integriert werden sollen. Die verwendeten Programmiersprachen waren Python (Flask) im Backend und JavaScript (React) im Frontend. Interessierte können sich gerne dieses Paper durchlesen.

Eine vorherige Masterarbeit behandelte die Einbindung eines MLOps-Systems in *SEDAR*. Machine Learning (ML) benutzt Algorithmen, um aus Trainingsdaten Muster zu erkennen und damit neue Daten analysieren zu können. So können ML-Systeme selbstständig lernen und sich adaptieren, ohne explizite Anweisungen zu erhalten. MLOps steht für ML und Operations (Ops) und ist ein Prozessverbesserungansatz für den Ablauf der Entwicklung von ML-Anwendungen. Hierzu besteht im System die Möglichkeit, ML-Modelle zu konfigurieren, in Jupyter Notebooks zu entwickeln und sie anschließend auch zu nutzen.

Automatisiertes maschinelles Lernen (AutoML) hat sich zu einem sehr wichtigen Forschungsthema entwickelt, bei dem maschinelle Lerntechniken Techniken [2]. Das Ziel von AutoML ist es, Menschen mit begrenztem Hintergrundwissen über maschinelles Lernen in die Lage zu versetzen, die Modelle des maschinellen Lernens Modelle einfach zu nutzen. Es gibt viele Arbeiten und Frameworks zur automatischen Modellauswahl und automatisierte Einstellung der Hyperparameter etc., automatisiertes Training und Evaluierung. Diese Arbeit soll an die bestehende Funktionalität zu MLOps anknüpfen und in *SEDAR* AutoML-Kapazitäten integrieren. Führende Cloud-Anbieter bewerben und bieten AutoML kostenpflichtig auf ihren Plattformen an; Azure ML kann in Rahmen dieses Projekts ausprobiert und als Inspiration verwendet werden.

Der Schwerpunkt der Arbeit kann von den Studierenden in Absprache mit den Betreuern gewählt werden. Mögliche Aufgaben sind im Folgenden beschrieben, sollen sich aber nicht strikt auf diese Ziele beschränken, sondern idealerweise den Prototyp als Ganzes weiterentwickeln:

Aufgaben:

- Recherche, Vergleich und Auswahl geeigneter **Open-Source** AutoML-Tools (siehe z.B. AutoKeras, Auto-Sklearn, H2O.ai, AutoGluon, TPOT, …)
- Erfassung und Vergleich der Funktionalität von Azure ML mit Open-Source-Frameworks
- Erweiterung des Metadatenmodells von *SEDAR* für AutoML
- Implementierung eines oder mehrerer AutoML-Tools in *SEDAR*
- Weitere Integration des *SEDAR*-JupyterHub
- Demonstration der Verwendbarkeit und Effektivität der implementierten Erweiterung an einem Use Case aus dem I^2DACH-Projekt

Ihr Profil:

- Informatik M.Sc. Student
- Interesse an interdisziplinärem Arbeiten
- Frontend-Programmierung mit JavaScript, vorzugsweise mit den Frameworks React und Angular
- Python OO-Programmierung

Interessiert? Fragen? Kontaktieren Sie uns!

Sayed Hoseini, M.Sc. – sayed.hoseini@hs-niederrhein.de

Prof. Dr. Christoph Quix – Christoph.quix@hs-niederrhein.de

Literatur

[1] C. Quix, R. Hai: Data Lake. In S. Sakr, A.Y. Zomaya (Eds.): Encyclopedia of Big Data Technologies. Springer 2019. https://doi.org/10.1007/978-3-319-63962-8_7-1

[2] Karmaker, Shubhra Kanti, et al. "Automl to date and beyond: Challenges and opportunities." *ACM Computing Surveys (CSUR)* 54.8 (2021): 1-36.