Design and Implementation of an ML-Based Data Cleaning Method for Structured Data

Thesis Supervisors:
Prof. Dr. Christian Scheideler
Prof. Dr. Axel Cyrille Ngonga Ngomo

Submitted By:
Sneha Hiremath

Abstract

Data science is growing widely and is used in most aspects of our daily lives. Data science domain experts analyze and make business decisions based on data from various sources. However, data in its raw form is of extremely poor quality, containing noise, errors like duplicate values, outliers, missing values and others. Domain experts spend the majority of their time cleaning data. This thesis introduces a novel data-cleaning method for structured data to aid naive users in designing ML Pipelines. There are several non-ML-based error detection methods dedicated to identifying specific error types. Few ML-based data cleaning methods concentrate on heterogeneous error types, but they prioritize performance over scalability. Additionally, they do not examine the effectiveness of the error detection method when used in ML pipelines. The adaptive data cleaning method in this thesis focuses on improving the data cleaning method's performance and scalability. An error detection method is developed by combining similar features from various ML-based error detection methods to form a feature pool. The implemented error detection method's feature sampler aids in improving the performance of error detection, while the active learning module minimizes the user involvement in the error detection process. Two settings are presented in this thesis, one to improve the efficiency of data cleaning and the other to reduce the model's runtime while maintaining acceptable performance. The data repair method is employed to repair the identified errors and provide an error-free dataset. The repaired dataset is then used in the ML model to evaluate the effectiveness of the implemented method. A comparison of Adaptive data cleaning and other existing methods shows that the former is more efficient and scalable than the latter.