REDS: Estimating the Performance of Error Detection Strategies Based on Dirtiness Profiles

Mohammad Mahdavi
Technische Universität Berlin
mahdavilahijani@tu-berlin.de

Ziawasch Abedjan
Technische Universität Berlin
abedjan@tu-berlin.de

Motivation
- Error detection is the task of finding wrong values
  - E.g., the red values in the table
- There are different error detection strategies
  - A rule violation detector with Kingdom → Lord [1]
  - A pattern violation detector with not-null [2]
  - An outlier detector [3]
  - ...
- Not all the strategies are always accurate [4]

Research Question
- Given a dataset and a set of error detection strategies, how can we estimate the performance of strategies without involving the user to evaluate them?
  - How can we automatically represent the dirtiness of datasets?
  - How can we identify the dirtiness similarity of datasets?
  - How can we leverage the dirtiness similarity of datasets to estimate the performance of strategies on a new dataset?

REDS estimates the performance of error detection strategies without any user labels via representing datasets by their dirtiness profile.

The Workflow of REDS

Dirtiness Profile
- Content features
  - Represent data domain
    - E.g., top keywords
- Structure features
  - Represent data type distribution
    - E.g., the fraction of numerical data values
- Quality features
  - Represent error type distribution
    - E.g., the normalized output size of an outlier detection strategy

Experimental Overview
- 11 Datasets
  - Hospital
  - Flights
  - Newspapers
  - Beer
  - Salaries
  - Address
  - Movies
  - Restaurants
  - Soccer
  - Tax
- Baseline
  - Maximum Entropy-Based Approach [5]
- Evaluation Measure
  - Mean Squared Error
- 4 Experiments
  - Effectiveness
  - Features
  - Regression Model
  - Repository Size

Experimental Results
- Maximum Entropy-Based [5]
- Unsupervised REDS
- Full REDS

References

Source Code
Our prototype is available online: https://github.com/bigdama/reds

Acknowledgement
This project has been supported by the German Research Foundation (DFG) under grant agreement 387872445.