CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks

Jeong Hoon Choi
USCID: 5023184813

DSCI-560
Data Science Professional Practicum

CleanML: Introduction & Related Work

CleanML

Data Cleaning
- Missing Values
- Outlier
- Duplicates
- Inconsistencies
- Mislabeled

ML Models
- LR
- KNN
- Decision Tree
- RF
- AdaBoost, XGBoost
- NB

- Framework to assessment performance of ML between data cleaning
- The results for each dataset, data cleaning technique, model, and scenario are stored in relational database R1, R2, and R3.

⇒ CleanML instance
- BD
  - model 1 trained w/ dirty dataset
  - model 2 " w/ cleaned dataset
  - t-test
- CD — model trained ;
  - test w/ dirty dataset
  - " w/ cleaned dataset

TABLE 2. Automatic Cleaning Methods

Error Type	Detection Method	Repair Method
Missing Values	Empty Entries	Deletion
		Mean_Mode, Mean_Dummy Median_Mode, Median_Dummy Mode_Mode, Mode_Dummy
		HoloClean
Outliers	SD	Mean, Median, Mode
	IQR	
	IF	HoloClean
Duplicates	Key Collision	Deletion
	ZeroER	
Inconsistencies	OpenRefine	Merge
Mislabels	cleanlab	cleanlab

CleanML: Database Schema

A. *The Three Relations*

TABLE 1. CleanML Schema. Keys are underlined.

R1 (Vanilla)

Dataset	Error Type	Detection	Repair	ML Model	Scenario	Flag

R2 (With Model Selection)

Dataset	Error Type	Detection	Repair	Scenario	Flag

R3 (With Model Selection and Cleaning Method Selection)

Dataset	Error Type	Scenario	Flag

- Flag: P, N, S (t-test results for 20 experiments)

- Relation R1: Vanilla

How does cleaning some type of error using a detection method and a repair method affect a ML model for a given datasets?

- Relation R2: Model Selection

*How does cleaning some type of error using a detection method and a repair method affect the **best ML model** for a given dataset?*

- Relation R3: Model Selection + Cleaning Method Selection

How does the best cleaning method affect the performance of the best model for a given dataset?

CleanML: Analyzed Database

TABLE 16. Summary of Empirical Findings for Single Error Types

Error Type	Impact on ML	Does the impact depend on			
		Datasets	Scenarios	Cleaning Algos	ML Algorithms
Duplicates	Varying (Mostly S & N)		No	Yes	No
Inconsistencies	Varying (Mostly S)		No	N.A.	No
Missing Values	Varying (Mostly P & S)	Yes	No	Yes	No
Mislabels	Varying (Mostly P & S)		Yes	N.A.	No (except Boosting)
Outliers	Varying (Mostly S)		No	Yes	No (except KNN)

- Missing Values: **imputation** >= deleting missing values
- Outliers: cleaning has insignificantly affected the performance
- Mislabels: cleaning has positive or insignificant impacts
- Inconsistencies: no significant impact (unlikely to have negative impact)
- Duplicates: cleaning is more likely to have insignificant or **negative** impacts than positive impacts

- Impact of cleaning on ML is inconsistent, depends on datasets

Better data cleaning than developing specific robust ML models

Question: What method did the author use to compare data cleaning and machine learning performance, and why?

Answer:

The author stored the machine learning performance data in a relational database for analysis.

This allowed for comparison of results across various combinations, and instead of simply comparing accuracy, the results were explained in terms of the given conditions P, N, and S.