CGI Internship: Refining an Old Database through Rule-based and Decision-based Cleaning

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Edward Kim

Fall, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Shangtong Zhang, Department of Computer Science

Abstract

While at Conseillers en Gestion et Informatique (CGI), my team ran into an issue of having to clean an old database by getting rid of certain entries from a poorly maintained database. My team was tasked with data cleansing, through two methods: decision-based cleaning (machine learning and AI algorithms) and rule-based cleaning (database queries and conditional restraints). Our solution utilized a variety of databases (SQLite3 and Postgres), AWS, sagemaker, Microsoft teams for communication, CSV files, gitlabs, and machine learning and AI algorithms. My task was to create scripts that inserted the contents of CSV files to SQLite3 or postgres databases, based on certain rules from the client. We achieved a certain confidence level by first using a rule-based process to filter the easier conditions, and then using the ML and AI algorithms to address the harder entries. Future work could be directed towards improving the confidence levels of the AI/ML algorithms, as well as improving ways the rule-based team can perform null detection based on SQL queries and other conditions.

1. Introduction

Imagine being a new employee tasked with writing queries to identify the people that fall under a certain criterion. Imagine the dread that arises when a 20+ -year-old database has multiple variations of a query as simple as "null."

Old databases with improper inputs are quite common, especially in the private sector. Improper values in databases lead to the wrong information being conveyed, which can lead to difficulty in search results or statistics depending on the level of discrepancy in the inputs.

This raises a question of what computer science-based solutions could be used to "clean" these old databases so that they are up to modern standards and follow a criterion or set of rules. In the project for CGI, we used AI/ML algorithms for decision-based cleaning and rule-based cleaning methods to bring old databases up to par with modern standards.

2. Related Works

Among the many discussions of cleansing large amounts of data using various methods, Ridzuan and Wan [1] point out the importance of cleansing big data in order to give accurate predictions for businesses and organizations, because miscalculations can be costly. They go into depth about the various processes in data cleansing, focusing on investigating traditional data cleansing for big data between 2013 and 2019.

This relates to my work because the old database my team and I worked on used various methods of data cleansing for a very large and very old big data database. However, my work differed in the sense that my team and I used a mixture of traditional and AI/ML algorithms to clean an old database.

My team used the benefits of traditional cleaning methods mentioned by Ridzuan and Wan [1], and we decided to also use ML algorithms to create a hybrid solution. Ihab and his team [2] bring up a good point that ML algorithms provide a decision based solution to the short comings of rule based cleaning by being able to have much more flexibility in data analysis compared to traditional methods such as "learning" the rules from the data over time without needing to keep track of a large number of rules, accumulate all the diverse set of signals and contexts for statistical analysis, and data profiling. In turn this allows for the ML algorithms to semi-automate cleaning exercises for my project.

Using a combination of these techniques provides a way for my team to thoroughly clean a database using both rule-based traditional methods and decision-based ML methods.

3. Process Design

The design of the system is focused on three components: the IDE for running the machine learning algorithm, the various databases used, and the machine learning algorithm itself.

3.1 AWS Sagemaker

We used AWS Sagemaker as an IDE that worked well with machine learning and connected to GitLab conveniently.

As an IDE, Sagemaker is similar to many of the other IDEs on the market; however, Sagemaker has the added benefit of being able to help improve model transparency by detecting statistical bias in ML. It can detect bias before or after training and it allows for different criteria of bias to detect. Also, Sagemaker is great for detecting imbalances during before, during and after data analysis.

3.2 Rule-Based Decision Algorithm

The rule-based algorithm my team used was straightforward in that we tried to clean the table based on only rules or set guidelines, expecting a certain output. For instance, we would determine that values that were greater than 20 digits were too large for the table and would be discarded, and any strings containing strange characters (usually Unicode characters) would be discarded. Once the dataset was cleaned based on these rules it would be passed along to the machine learning algorithm for clustering in order to target specific inputs that fell outside of the scope of the rule-based cleaning.

3.3 Machine Learning Algorithm

The type of ML algorithm that was used on my team was unsupervised statistical learning. With unsupervised learning, the ML algorithm has no defined outputs; in other words, it does not know the output. It creates its own output based on a few principles such as clustering, in which the ML algorithm discovers the inherent groupings of the data. One such example can be seen below where a ML algorithm will find certain properties of the inputs and then cluster or group them into different categories. The example in Figure 1 below clusters them based on shapes.

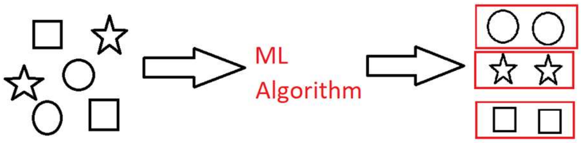

Figure 1: Unsupervised ML Example

We used this unsupervised learning approach method to essentially go through the database and cluster certain responses based on the type of data entry. For instance, in order to detect "null" entries, the ML algorithm might cluster various possible entries such as "-1," "NaN," "null," "NULL," etc. Once the ML algorithm clusters the dataset, analysts inspect the clusters and on some confidence interval determine how accurate the clustering is.

4 Results

My code was used mostly for the rule-based cleaning portion of the team for which I had to create scripts that import data from a csv file and help filter out the dataset, inserting it into a particular database for testing purposes. My scripts should help those on the rule-based decision-making team with their scripts whenever a csv file is involved and they want to run it against certain rules and criteria. Once all of models are improved on and everything is fully implemented, the ideal scenario is that CGI can clean the large government database using this combination of techniques.

5 Conclusion

The objective of this project was to clean an old database using two main methods: rule based and decision-based algorithms. The rule-based algorithm focused on using SQL queries that ensured that rules or set guidelines, expecting a certain output. The decision-based algorithm focused on unsupervised learning and uses clustering to clean the database within a certain confidence interval. My work on the rule-based portion of this project helped other members of the team by providing scripts that apply these rules to CSV files, and inserting them into SQLite3 or POSTGRES database tables. Even though the overall project is not close to being finished, using rule-based and decision-based cleaning has proven to be a good strategy in cleaning old databases. Hopefully, CGI is closer to its goal with some of the work I have provided during my time there.

6 Future Work

Currently, the team is improving both the rule-based and decision-based portions of the project. In terms of the rule-based portion of the project, the team is still trying to determine the best rules to use in order to clean the database keep track. For the decision-based portion, the team still needs to improve its ML model by finding better datasets to establish a better confidence interval. I expect the people on my team to finish this project in a couple years.

References

[1] Ridzuan, F., & Wan Zainon, W. M. (2019). A review on data cleansing methods for Big Data. *Procedia Computer Science*, *161*, 731–738. https://doi.org/10.1016/j.procs.2019.11.177

[2] Ihab, F. Ilyas University of Waterloo, Ilyas, I., Waterloo, U., Wisconsin, T., Rekatsinas, T., Wisconsin, U., . . . Metrics, O. (2022, September 01). Machine Learning and data cleaning: Which serves the other? Retrieved October 22, 2022, from https://dl.acm.org/doi/10.1145/3506712