Udata - A machine learning approach to Data Cleansing

Hara Acharya, Nitin Dutt

Abstract

Data Cleansing is a typical approach used for Machine Learning. Data only presents true value when reflecting complete accuracy. Humans make mistakes, which is how inaccurate data usually finds itself on computer logs. We propose AI programming could provide a course correction when human foibles lead to data inaccuracies. AI's ability to perform a multitude of tasks continues onward. We propose below two methods which can help cleanse the data.

- Iterative methods like use of stochastic gradient descent (SGD) for correcting, updating, repairing, and improving data before model creation. We are still studying SGD to come up with a solution.
- Neuro-fuzzy modeling to produce a unique adaptive framework for entity resolution, which automatically learns and adapts to the specific notion of similarity at a meta-level.

While we are still studying and trying to measure performance implications of iterative methods for Data cleansing, we have already built the theoretical framework using neuro-fuzzy modeling, which removes the repetitive task of hard coding a program based on user defined rules. Cleansing Data with user defined rules is being used currently for data cleansing.

We think this neuro-fuzzy modelling based framework can be utilized in the production of an intelligent tool to increase the quality and accuracy of data.

Introduction:

When data are gathered from distributed sources, differences between tuples in database or csv files are generally caused by four categories of problems in data, namely:

The data are-

- Incomplete.
- Incorrect.
- Incomprehensible.
- Inconsistent.

Some examples of the discrepancies are spelling errors; abbreviations; missing fields; inconsistent formats; invalid, wrong, or unknown codes; word transposition; Very interestingly, the causes of discrepancies are quite similar to what has to be fixed in data cleaning and preprocessing in data warehouse or DataLake.

For example, in the extraction, transformation, and load (ETL) process of a data warehouse, it is essential to detect and fix these problems in dirty data.

Elimination of fuzzy duplicates should be performed as one of the last stages of the data cleaning process.

Proposal for the solution:

Removing inaccurate or unreliable data serves as the first step to fixing things. The next step involves inputting the correct data to replace the inaccuracies.

How to think about solving this problem using AI.

The problem has 2 parts.

- Understanding the data.

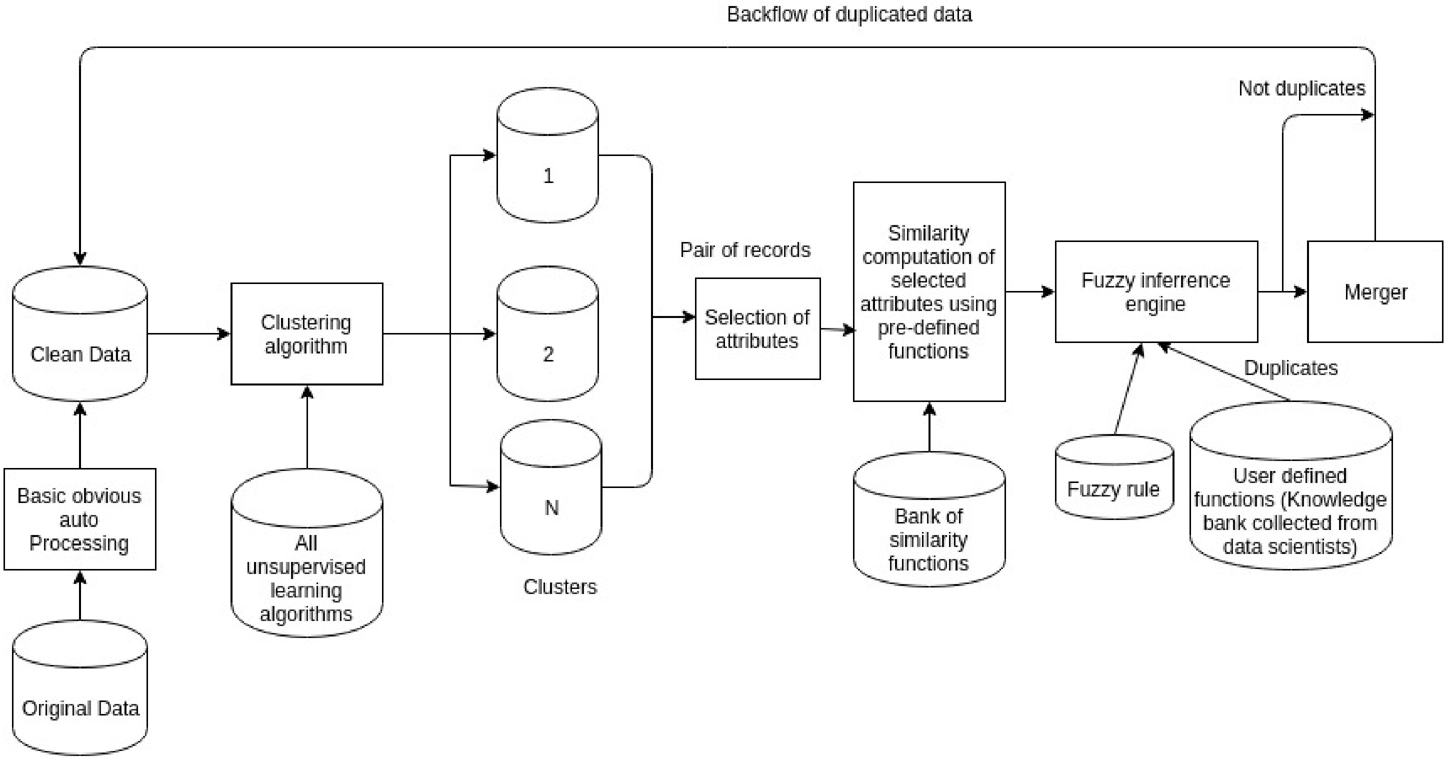
And

- Discovering an errors/discrepancies in the Data set.
- Removing duplicate rows/tuples in the data set. **(Not applicable to sensor data)**
- Once error for example “out of boundary values in numeric sensor data" are discovered, replace them with accurate values.

Example of various discrepancies:

Discrepancy type	Name	Address	Phone Number	ID Number	Gender
	Nitin Dutt	Remington Drive 160	615 5544	553066	Male
Spelling Error	Nitin Dut	Remington Drive 160	615 5544	553066	Male
Abbreviation	N Dutt	Remington Dr 160	615 5544	553066	Male
Missing Field	Nitin Dutt		615 5544	553066	
Different format	Nitin Dutt	Remington Drive 160	+1-650-615-5544	553066	Male
Word Transposition	Dutt, Nitin	Remington Drive 160	615 5544	553066	Male

Data cleansing algorithm implementation for removing duplicate data after basic cleansing:


Backflow of duplicated data
Not duplicates
Clean Data
Clustering algorithm
1
2
N
Clusters
Pair of records
Selection of attributes
Similarity computation of selected attributes using pre-defined functions
Fuzzy inferrence engine
Merger
Duplicates
Basic obvious auto Processing
Original Data
All unsupervised learning algorithms
Bank of similarity functions
Fuzzy rule
User defined functions (Knowledge bank collected from data scientists)

step-1: On original Data, we will have a set of rules to do auto/obvious cleansing with set of predefined rules.
Step-2: On the cleaner data, we will use clustering which will take care of below cleansing.

- **Data standardization/attribute correction.**
- **Duplicate matching.**
- **Outlier detection.**

How clustering can help in above cleansing.

As clustering is an unsupervised model, we can’t calculate the confusion matrix to evaluate the model. So we will follow the below method.

- Determine how close each data/column data within each cluster is to every other column data.(the intra cluster distance)
- How close each cluster data is to other clusters. (the inter-cluster distance)
- And compare the 2 distances.

Models that produce smaller intra-cluster distance and large inter-cluster distance evaluate favourably.

We also have to do below things as part of cleansing. But we have to use classification using neural networks to solve the below 2 problems.

- **Validation rules generation**
- **Missing attribute prediction**