Smart Data Cleansing & Predicting

PRABHJOT SINGH
CSE-AIML

CHANDIGARH UNIVERSITY

Chandigarh,Punjab,India

20BCS6895@cuchd.in

GAURAV SHARMA
CSE-AIML

CHANDIGARH UNIVERSITY

Chandigarh,Punjab,India

20BCS6879@cuchd.in

RAVI

CSE-AIMl

CHANDIGARH UNIVERSITY

Chandigarh,Punjab,India

21BCS8808@cuchd.in

I. ACKNOWLEDGEMENT:

I would like to express my special thanks of gratitude to my teacher Dr. Amit Garg who gave me the golden opportunity to do this wonderful project on the topic data cleaning, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them.
Secondly I would also like to thank my team members and friends who helped me a lot in finalizing this project within the limited time frame.

II. INTRODUCTION:

Data cleaning is the process of fixing or removing incorrect, despoiled, incorrectly formatted, replacement, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.
Step 1: Remove duplicate or irrelevant observations
Step 2: Fix structural errors
Step 3: Filter unwanted outliers
Step 4: Handle missing data

Data cleaning is the initial stage of any machine learning project and is one of the most critical processes in data analysis. Data cleaning entails a slew of procedures that, once done, make the data ready for analysis. Given its significance in numerous fields, there is a growing interest in the development of efficient and effective data cleaning frameworks. In this survey, some of the most recent advancements of data cleaning approaches are examined for their effectiveness and the future research directions are suggested to close the gap in each of the methods.

As a data scientist, one of the biggest struggles is cleansing the data before one can actually dive into it to get some meaningful insights. Data cleaning is one of the most important steps which should never be ignored. If the data is not cleaned thoroughly, the accuracy of your model stands on shaky grounds Poor quality of

data leads to biased results with low accuracy high error percentages thus, it is important to clean the data thoroughly before fitting a model to it. As a data scientist, it is important to understand that all the data provided to us may not be useful and hence we must know the ways to treat them. What is data cleaning – Removing null records, dropping unnecessary columns, treating missing values, rectifying junk values or otherwise called outliers, restructuring the data to modify it to a more readable format, etc is known as data cleaning.

Content of Dataset

This dataset contains about 10 years of daily weather observations from many locations across Australia.

RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

Context

Predict next-day rain by training classification models on the target variable RainTomorrow.

Machine learning concpets used :

Classification
Binary Classification

III. LITERATURE REVIEW:

- ***Review 1***

Data keeps increasing, but the quality of the data is decreasing as many of the data collected is dirty.

Cleansing approaches are available to solve this issue but data cleansing remains as a challenge in order to cope with some of the approaches are not suitable for big data as it has a significant amount of data that despite the availability of existing frameworks to address data cleansing for big data expert in undeniable as an expert is needed to verify and validate the data before it can undergo an analysis process.

- ***Review 2***

Data cleansing has become a major activity performed by most organizations that have data warehouses. Every organization needs quality data to improve on its services it renders to its customers. In view of this a thorough review of approaches and papers in that regard are discussed and their limitations also stated. This is to help future development and research directions in the area of data cleansing. The papers reviewed in this report looked at critical aspects of data cleansing and the various types of data that could be cleansed. Several algorithms have been proposed in the various works discussed.

- ***Review 3***

Data cleaning is very necessary part of data mining. From the above study we can see that there are different types of problems in data cleaning. Data cleaning methods and approaches depend upon the type of data which we want to clean and according to that we apply particular methods. This paper also present a comparison of data cleaning tools and determines the best tool. Each tool has its own specific features and depending upon the data we can use the tool to clean data. In future work we can check other functionality of these tools and suggest own.

- ***Review 4***

A number of authors have proposed a solution to address data cleansing problems. Traditional data cleansing methods is called traditional because it is not Meanwhile, some of the methods are designed specifically for big data like Cleanix, SCARE, KATARA, These methods are developed

to address the issue arises when dealing with big data during the cleansing process.

IV. WHAT IS YOUR IDEA?

The idea of this project is to use past weather forecasting data to predict the visualization of future weather forecasting prediction to help people (like farmers , for saving their crops from danger and many more). First unprocessed data is take and all the incorrect values are changed and raw data is maded fit for taking it as input data then all the data is processed and weather prediction is maded

- **Resolve Empty Values:**

  Data cleansing tools search and fill up those missing values to complete the data set and evade gaps in information.

- **Standardised Cleaning Process**:

  The data cleaning process must be standardized as per repeated testing and methods which proved to produce quality data, which later on helps in easy replication and consistency.

V. Conclusion:

A data cleaning tool will alter most aspects of an entity's general data cleansing program, but this data cleaning tool is just a part of an ongoing remedy for data cleaning. The outline of the data cleaning steps are as below:

- **Identify critical data fields**:

  The primary step is to identify which types of data fields are crucial for the intended project.

- **Data collection**:

  The data contained in the short-listed data fields is collected, classified and organised.

- **Discard duplicate values**:

  Duplicate figures are recognised, eliminated, and inaccuracies are resolved.

VI. REFERENCES

1.The Fifth Information Systems International Conference 2019
A Review on Data Cleansing Methods for Big Data
Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon

2.
A Review of Data Cleansing Concepts – Achievable
Goals and Limitations
AN OVERVIEW STUDY ON DATA CLEANING, ITS TYPES
AND ITS METHODS FOR DATA MINING

3.
S.LakshmiMphil Research scholar-VISTAS
Dr.S.v Prof & Head Dept of computer application VISTAS
lakshmi_lk@yahoo.com,

4.

Babar, S.; Mahalle, P.; Stango, A.; Prasad, N.; Prasad, R. (2010): Proposed security model and threat taxonomy for the Internet of Things (IoT). International Conference on Network Security and Applications, pp. 420-429.

5.
Berti-Equille, L.; Dasu, T.; Srivastava, D. (2011): Discovery of complex glitch patterns: a novel approach to quantitative data cleaning. IEEE International Conference on Data Engineering, pp. 733-744.

6.
Bertossi, L.; Kolahi, S.; Lakshmanan, L. V. S. (2013): Data cleaning and query answering with matching dependencies and matching functions. Theory of Computing Systems, vol. 52, no. 3, pp. 441-482.