Data Cleaning and Visualization using Machine Learning

C.J.Subhaswini
Final year Student, Information Technology, Loyola-ICAM College of Engineering and Technology, Chennai, India.
Email:cjsubbu@gmail.com
A.Sharmila
Final year Student, Information Technology , Loyola-ICAM College of Engineering and Technology, Chennai, India.
Email:asharmi98@gmail.com
G.Shobana
Assistant Professor, Information Technology, Loyola-ICAM College of Engineering and Technology, Chennai, India.
Email:shobana.g@licet.ac.in
Raina Miriam Jose
Final year Student, . Information Technology, Loyola-ICAM College of Engineering and Technology, Chennai, India.
Email:raina.joes2015@gmail.com

ABSTRACT

Data cleaning is a time taking and challenging process which is a basic necessity before moving into data analysis. This paper proposes an application which performs data cleaning which is a prerequisite for data analysis and then provides a visual representation of the cleansed data. This application will take in the structured dataset that contains both textual and numerical data which are then processed using machine learning algorithms to obtain a cleansed dataset. This process undergoes a series of steps to clean the data so as to acquire efficient results.

Keywords - **data cleaning, data visualization, machine learning, processing, analysis**

I. INTRODUCTION

Many organisations use datasets for predictive analysis and an important concern in these cases is data quality. Using noisy data can hamper with the correctness of analysis. The common errors are missing values, duplicates and other errors. These errors need to be corrected for reliable decisions and analytics. The users must be knowing the effects of using the noisy data before proceeding with the cleaning process. Noise removal can improve model performance, due to the fact that noises may disturb the discovery of important information. Therefore, this step of pre-processing is very crucial for the efficient working of the processing steps that follow it.

One of the crucial and widely appreciated applications of artificial intelligence is machine learning. Machine learning is used to allow the system to learn automatically without human assistance. It will help in analysing huge datasets with a large number of data fields. Machine learning is considered as a boon to the growth of most of the industries as it helps in making the work much simpler when compared to other technologies that are existing. With the data provided by the system after implementing the machine learning algorithms, organizations are able to work more effectively and acquires profit over their competitors. The system that uses machine learning will be able to predict how the structure looks like and just adjust the data according to the structure. Automation can be made simpler since machine learning uses an iterative approach to learn from data.

The major challenge with machine learning is to deal with large data sources in applying machine learning models for cleaning process. Data cleaning process is carried by taking in huge datasets which are then checked for the possible errors by using machine learning algorithms. The other challenges that include are avoiding learning process from noisy data, avoiding building a prejudiced model, not giving reasons for compromising with the quality of the data. A huge amount of time is spent in cleaning the dataset and creating an error-free dataset when it comes to utilizing machine learning data. The best practices that are used for data cleaning using machine learning are filling missing values, removing unnecessary rows, reducing the size of the data and implementing a good quality plan.

The success of machine learning applications depends on the amount of good quality data that is given to it. But this process of cleaning may not be considered as a main area in processing and most often they aren't mentioned but it is really critical when comes to providing predictions based on the data. The system that uses powerful algorithms to process the noisy data can yield bad results if irrelevant or wrong training of data is given. Machine learning comes into picture when the whole process of splitting the corrupted data from the good data is done in a large amount of time. We use ML algorithms to find out the different patterns in the data and group it by itself into clean and noisy data which will help in reducing a lot of time.

In this paper, we will focus on removing the columns and rows with less information, identifying the numerical values, predicting and filling in missing values and detect outliers which hamper with data analysis. We propose a system that simplifies the process for the user and allows for better processing.

In summary, Machine learning for data cleaning might be the only way to provide complete and trustworthy data sets for effective analytics.

II. ARCHITECTURE

In this devised model, the idea projected is to use the data cleaning application to process the raw dataset containing both textual and numerical data and convert it into a clean dataset which can be further used for data analysis.
Initially, users must upload the dataset which they want to clean into the application. Users can choose the operations that they want to perform on their dataset from the modules provided. This application performs a series of operations which includes removing columns with less information or no information, removing unnecessary rows, identifying the numerical values, filling in the missing fields and identifying the outliers.

Fig 1: System architecture

In the first module, some columns may contain less information or no information at all, which makes it hard to rely on such columns for analysis and so such columns can be removed provided that they don't cause significant damage to the process.

In the second module, some rows may contain empty fields which will again tamper with the proper pre-processing of the dataset. Hence such values are identified and removed.
In the third module, the dataset will contain categorical features ranging from numerical to non-numerical values. This application requires only numerical data which is used for analysis and prediction. So, the fields containing numeric values are identified.

In the fourth module, we deal with missing values which occur for a multitude of reasons —ranging from human errors during data entry, incorrect sensor readings to software bugs in the data processing pipeline. It is probably the most widespread source of errors and the reason for most of the exception-handling. If you try to remove them, you might reduce the amount of data you have available dramatically. So, these fields need to be filled in with appropriate values.

In the fifth module, we deal with outliers which are those data points that are really far from the rest of your data points. Mathematically, an outlier is usually defined as an observation more than three standard deviations from the mean. They can show up due to errors in data entry or measurement, or just because there's a variation in the population. Identifying and handling outliers is an important part of data cleaning.

III. RELATED WORK

There are various existing studies carried out by many authors on data cleaning and visualization. The aim of this approach [11] was to generate an abstract clean instance which is the perfect approximation of all feasible concrete clean instances. The Abstract Interpretation framework was applied to achieve this aim. The technique used had the following four phases: Clustering, Abstraction, Application of MD and Sound Query Answering. In the clustering step, a similarity metric was used to form a set of clusters by grouping the values. In the abstraction phase, abstract domains were applied aiming at replacing concrete values by appropriate properties in which the user is interested in and the Abstract Interpretation theory was followed. An abstract clean instance was created which is the perfect approximation of all possible concrete clean instances. It reduced the computational complexity of query processing and the space requirement. For cleaning dirty databases, the domains of intervals were chosen as an abstract domain for numerical values and bricks was chosen as the appropriate abstract domains for numerical and string values respectively. In the third phase, the application of MDs and matching functions to clean dirty data in the abstract domain had been discussed. Next the matching function for Numerical abstract values were defined. It was done by lattice least upper bound. The matching functions were defined in terms of least upper bound of the corresponding lattices for sign and parity abstract domains. The preciseness of the clean results was improved. In the final step, abstraction was applied to the query's variables or parameters to obtain a sound query answer and to compare it with the obtained abstract clean instance. The final result was an over-approximation of the concrete results which provided the users with even more information.

An overview of qualitative data cleaning with error detection and error repairing approaches were discussed in [10]. Rule-based data cleaning techniques were focussed on where errors like duplication, inconsistency and

missing values were dealt with. It also described a statistical perspective on qualitative data cleaning using Machine Learning techniques.

Bid data Landscapes [8] deals with three methods for the process namely Pixel level, Glyph level and details on demand. It uses appropriate MDS algorithms to reduce the dataset dimensions and prepare it for visualization. In the pixel level, one can choose between categorical and ordered color-coding so that a particular cluster or feature value can be visualized. In the glyph level, star plots or flower glyphs can be used according to the dataset in hand as both have varied advantages. Glyph visualizations help one recognize patterns in the data and identify divergent feature values. In details on demand, overlays like Tooltips and configuration overlay have been used. The data is accepted using a JSON-API which describes various attributes. AWS was used for computing machine learning algorithms and MDS algorithms. The system was implemented using JavaScript and Angular as the application platform. D3 library was used to draw on a canvas component for data visualization.

IV. METHODOLOGY USED

The application is developed using Angular. It is a dynamic framework for developing web application. Single page applications can be effectively created in a clean manner using angular. Data Cleansing modules such as removing columns with less information, removing unnecessary rows, identifying numerical values, filling in missing fields and handling the outliers are performed using the python code with the help of machine learning algorithms. Each and every module uses the python code to perform the data cleaning process step by step. The final output will give the cleansed dataset.

The web application and the python code are integrated by invoking a service file in the angular project. Web service contains a set of code that can be invoked remotely with the help of HTTP. The Web service can be activated using HTTP requests. It helps in exposing the functionality of the existing code over the network.
The process flow for this system are as follows, the dataset uploaded by the user in the application is transferred to python which runs in a particular port in the network through the service file. The service file uses HTTP to transfer the dataset to the python code and returns the cleansed dataset through HTTP to the application. This application will be used by the end users who wants to clean the dataset which they want to use for analysis or prediction.

V. EXPERIMENTAL RESULTS

The system works as explained below
The user can click on the 'Get Started' button provided and then select the operations they wish to perform on their dataset from the list of operations provided. The user can then upload the dataset into the application and click on the Upload button to start the cleaning process. Initially the original dataset is displayed and the dataset after operation 1 is displayed with an Export option to download the cleansed dataset. The selected operations are performed successively and finally the completely cleansed dataset is obtained.

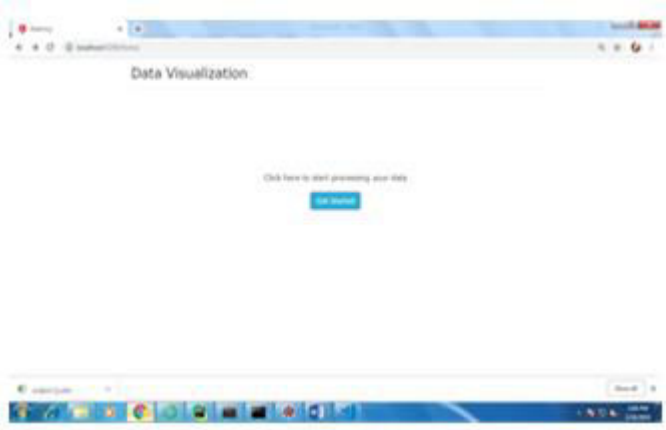

Fig 2: Start of the data cleaning process

Fig 3: Users can choose the operations they want to perform on their dataset

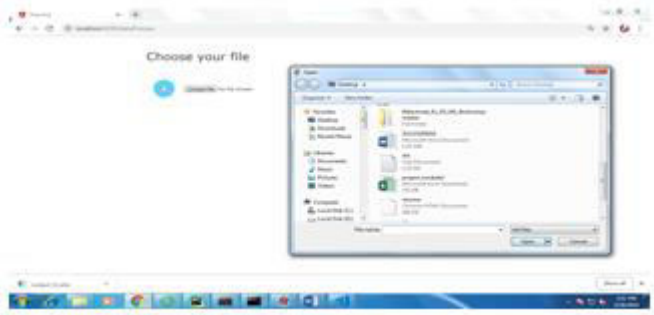

Fig 4: Upload the dataset

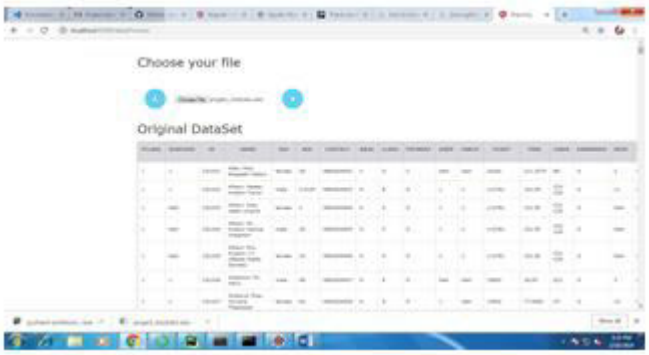

Fig 5: View of the dataset uploaded

Fig 6: View of the dataset after processing the first module

Fig 7: Users can download the cleansed dataset

VI. FUTURE WORK

The current system performs the five operations as mentioned above to clean the dataset. We propose to include more operations to clean the other commonly found errors in datasets to provide more accurately cleansed datasets for richer analysis. And also processing with the cleansed dataset to provide analysis with the data will be continued in our future work. Data pre-processing consists of four processes which includes data cleaning, data integration, data transformation and data reduction. Our system performs only one part of the data pre-processing. All the other pre-processing steps will be implemented in our future work.

VII. CONCLUSION

The Data Cleaning is considered as a main challenge in big data era since detecting and repairing the raw data manually is a hectic process and takes a lot of time. Hence, we have proposed a system which takes in the raw datasets into the application which are then preprocessed to clean up all the dirty data using machine learning algorithms. Finally, the cleansed data is visualized to the users after all the preprocessing is done. This system saves a lot of time since manual cleaning can be avoided. This serves as an effective purpose for the users who wants to clean huge datasets.

REFERENCES

[1] TextTile: An Interactive Visualization Tool for Seamless Exploratory, Analysis of Structured Data and Unstructured Text, Cristian Felix, Anshul Vikram Pandey, and Enrico Bertini, Member, *IEEE*

[2] Efficient Outlier Detection for High-Dimensional Data,Huawen Liu, Member, IEEE, Xuelong Li, Fellow, *IEEE*, Jiuyong Li, Member, IEEE,and Shichao Zhang, Senior Member, *IEEE*

[3] A Statistical Direct Volume Rendering Framework for Visualization of Uncertain Data,Elham Sakhaee, Alireza Entezari, Senior Member, *IEEE*

[4] An Automatic Big Data Visualization Framework,Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li, IEEE

[5] Sequence Synopsis: Optimize Visual Summary of Temporal Event Data Yuanzhe Chen, Panpan Xu and Liu Ren, *IEEE*.

[6] C. Ahlberg and B. Shneiderman, "Visual Information Seeking:Tight Coupling of Dynamic Query Filters with Starfield Displays," in Proceedings of *the ACM Conference on Human Factors in Computing Systems,* New York, NY, USA, 1994, pp.313–317.

[7] F. Beck, S. Koch, and D. Weiskopf, "Visual Analysis and Dissemination of Scientific Literature Collections with SurVis," *IEEE Transactions on Visualization and Computer Graphics,* vol.22, no. 1, pp. 180–189, Jan. 2016.

[8] Big data Landscapes: Improving the visualisation of Machine-Learning based Clustering algorithms, Deitrich Kammer, Mandy Kreck, Thomas Grunder and Rainer Groh, *IEEE*.

[9] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-driven documents," *IEEE Transactions on Visualization and Computer Graphics,* vol. 17, no. 12, pp. 2301–2309, Dec. 2011.

[10] Keshif: Rapid and Expressive Tabular Data Exploration for Novices, Mehmet Adil Yalçın ; Niklas Elmqvist ; Benjamin B. Bederson, *IEEE Transactions on Visualization and Computer Graphics* (Volume: 24 , Issue: 8 , Aug. 1 2018)

[11] Data Cleaning: Overview and Emerging Challenges ,Xu Chu, Ihab F. Ilyas, Sanjay Krishnan and Jiannan Wang, *IEEE*.

[12] Data Cleaning: An Abstraction-based approach, Dileep kumar koshley and Raju Hadler, *IEEE*.

[13] Interactive visualisation of large datasets, Parke Godfrey, Jarek Gryz and Pioter Lasek, *IEEE*.

[14] Machine Learning to Data Management: A Round Trip , Laure Berti-Equille , Angela Bonifati and Tova Milo, *IEEE*.