The Novel Method for DataPreprocessing CLI

[1]Chithra Y, [2]Prathibha Kiran. D and [3]P B Manoj
[1,2,3]ECE Department, AMCEC, (Affiliated of VTU), Bangalore, India.
[1]chithray03@gmail.com, [2]prathibha.swamy87@gmail.com, [3]manoj0110@gmail.com

Article Info
Jenitta J and Swetha Rani L (eds.), *International Conference on VLSI, Communications and Computer Communication*, Advances in Intelligent Systems and Technologies,
Doi: https://doi.org/10.53759/aist/978-9914-9946-1-2_21

Abstract— Data preprocessing is the first step in machine learning to ensure data quality and extract useful information from datasets. Derived data after data processing is used for model training and has a direct impact on model efficiency. If there is no relevant and dispensable information in the dataset, it will be removed from the dataset to ensure data quality. Data pre-processing includes description of data, null value handling, categorical value coding, normalization, transformation, extraction and selection of various features.

Keywords— Data Preprocessing, Dataset, Machine Learning**.**

I. INTRODUCTION

Machine learning is a domain of artificial intelligence which focuses on data. In machine learning the data description is a first step in which we encode, transform the data to train the model. Data can be in different types like images, videos and audios etc.

Data needs to be preprocessed, [1] an important phase of data mining. User recognition and data cleansing are the methods in data preprocessing. The purpose of data cleansing is unrelated data. Thiscurrent study continues to preprocess the data methods that include data cleansing, data integration, data transformation and data reduction. There are many different techniques available for data sanitization, but they do exist. Accurate metrics for some data collection issues and user identification of the data preprocessing isvery important.

Data preprocessing is used to clean up data in a given pattern or direction. Discovery identifies the techniques further used to discover user navigation patterns. After processing, proceed to pattern analysis. This gets only the relevant patterns and removes the unrelated patterns. [14] Data mining is one such method for finding these patterns in an abundant amount of raw data. Dataset is one which affects the type and efficiency of data processing. Finding a good data source, also improves quality patterns and algorithms. Indata preprocessing and data collection differ not only in the type of data available, but there are also source websites, source sizes and approaches will be processed.

The purpose of data preprocessing is to provide something that is reliable, structured, and integrated. Pattern identification, Statistical studies, clustering, classification, and much more are used to discover rules and patterns. The following knowledge has been discovered in the form of visual elements like charts, graphs, rules, etc. and can characterize, compare, predict, or classify data from a data set.

To preprocess data, [15] we need a dataset, dataset is a collection of data. Dataset which is represented in the form of rows and columns. Preparing data for training a model is a very important step, because algorithms cannot work accurately on raw data. Proper dataset is required to solve the issues and to arrive at decisions. Before applying the dataset to any machine learning model which needs to be converted in a way that the algorithm understands the dataset.

The main aim of the data preprocessing is to ensure the quality of data to train the machine learning model, completeness of data set and consistency of the dataset to train the model. So, implement a command line interface (CLI) which will preprocess a dataset and save time. Command line interface is very easy to use, as by providing the commands in command prompt the data can bepreprocessed. It provides the ease of access to the dataset, to fetch the data which has to be preprocessed and also to access the rows and columns of the dataset which is already preprocessed.

II. RELATED WORKS

When data is insufficient or contains unrelated and irrelevant information, machine learning is essential. Algorithms are not always right and can be difficult to understand to find the result, or they don't help at all. Therefore, preprocessing the data is a mandatory step in the machine learning process. This requires a preprocessing step to solve various types of problems, including dispensable data, noise induced data, lost data values, etc. All learning algorithms heavily rely on the results at this point. This is the final training set. [2]

Data mining algorithms search raw data sets for meaningful patterns. The data mining process requires a great deal of computing power for large data sets. Reducing the size can effectively reduce this cost. This is a pre-processing step where a dimension is removed from a given data set before it is passed to a miningalgorithm. This paper explains how it

is often possible to reduce size with minimal loss of information. A clear classification of dimensionality reduction is described and techniques of dimensionality reduction are theoretically presented.[3]

The problem of multiple consolidation of common entity information databases is common in various fields of business-related activities, both in government and private sectors. The problem being studied here is termed as *merge/purge problem* and requires excess of time and money to solve both in terms of scale and precision. Big data stores often have many duplicate entries of information about similar entities that are difficult to put together without an intelligent "equilibrium theory" that identifies equivalent entries through a process complex, domain-dependent matching. Wehave developed a system to accomplish this task of data cleansing
and demonstrated its use for cleansing a list of leads in a direct marketing application. Our results for statistically generated data are accurate and efficient when the data is processed multiple times and different sort keys are used on each pass. Combining the results of individual runs using a bridge of independent results yields much more accurate results at lower cost. The system provides a rules module that is easy to program and very effective in finding identical data, especially in environments with large amounts of data. The report of a successful database implementation under real conditions clearly confirms earlier results. for statistically generated data.[4]

Statistics and Data Mining are two domains commonly used in data analysis and knowledge discovery. While statistics involves applied mathematics, data mining is a multidisciplinary domain thatdeveloped from computer science, but both are used for the same purpose. There are many approaches that the two fields share in common. But some approaches used in statistics can decrease the workload of a data miner. The growth of data mining has been tremendous in the last decade. Its application has increased with the growth of generations of data. More and more research is being done in the field of databases with the help of data mining. This is because data mining can be used in advanced data analysis and has the potential to extract indispensable knowledge from massive data sets. It has become a new science and technology to meet these needs. Data mining is often referred to as solving a problem by analyzing data that already exists in a database. In addition to mining structured and numerical data stored in datasets, there are now an increasing number of interested parties who have experiencemining unstructured and non-numerical data such as text and web.[5]

The study and analysis of this data can help in the organization of services ranging from website customization, system improvement, website modification to business intelligence to determine the characteristics of use. The study and analysis of this data can help in the organization of services ranging from website customization, system improvement, website modification to business intelligence to determine the characteristics of use. Web mining is the applicationof data mining techniques to extract knowledge from web data - including web documents, hyperlinks between documents, website usage logs, etc. Web crawling is broadly divided into crawl web content, explore web usage, and web structure.[6]

III. METHODOLOGY

Titanic dataset: Titanic dataset is used to preprocess. It is one of themost popular datasets used to understand the basics of machine learning. [7] It contains information about all the passengers aboard the RMS Titanic, which unfortunately sank in the depths of the Atlantic Ocean. This dataset (Fig 1) can be used to predict whether a particular passenger survived or not.

Variable	Definition
survival	Survival {0 = No, 1 = Yes}
pclass	Ticket class {1 = Class A, 2 = Class B, 3 = Class C}
sex	Sex
Age	Age in years
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation {C = Cherbourg, Q = Queenstown, S = Southampton, O= Others}

Fig 1. Overview of Dataset

The dataset is preprocessed with different steps, The preprocessing steps are:

Input the data set: [8] Here titanic dataset is used. “There are several types of Machine Learning such as Supervised learning, Unsupervised learning etc. Here, we are writing python scripts to make a preprocessed dataset for performing supervised learning.” “Supervised learning consists of mapping input data (independent variables) to known targets (dependent variable), which humanshave provided. Predicting house prices is a good example.”

Data description: Here we implement the functionality which will enable the users to describe the dataset properties like mean, max, standard deviation etc. To show main statistical details like mean, maximum, minimum, percentiles etc., datatype of columns of datasetuses standard library like pandas, numpy.

Handling the null values: [9] The next step in data preprocessing is to deal with missing data in the datasets. If your dataset contains some missing data, it can cause big problems for your machine learning model. Since there are sometimes missing values in the data, it's important to handle them properly. Handling missing values is also known as data imputation.is milestone aims to rid the dataset of all null values.

Encoding Categorical Data: [10] Categorical data is data which can be divided into categories. Machine learning models work solely with math and numbers, but if your data is classified in a categorical way,it may cause trouble while building the model. It is necessary to convert these categorical variables into numbers in order to analyze them statistically. The goal of this milestone is to assign numerical values to all the categorical columns in the data.

Feature Scaling: [11] Feature scaling is a method of standardizing the range of an independent variable or column of data. To cope with large differences in column size, the data are normalized. If feature scaling is not performed, the machine learning algorithm usually weights larger values more heavily and considers smaller values as the lower values, regardless of the unit of the values. To avoid this, feature scaling is performed.

There are two main ways of scaling features:
Normalizing: [12] It is a technique to make the values in a numeric column on the dataset more comparable, you can simply change the scale used to display the data. This will not affect the actual values in the column, but it will make comparisons between the values easier.

Standardizing: [13] It is a technique where the values are centered around the mean and have a uniform standard deviation. The mean of the attribute decreases and the resulting distribution has a uniformstandard deviation.

Once the preprocessing is complete, we can start downloading the preprocessed data.

IV. RESULTS

Table1. Preprocessed DATASET

Passenger id	pclass	Name	age	sibsp	parch	fare
1	3	Braund, Mr. Owen Harris	22	1	0	7.25
2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	38	1	0	71.2833
	3	Heikkinen, Miss. Laina	26	0	0	7.925
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35	1	0	53.1
5	3	Allen, Mr. William Henry	35	0	0	8.05

As we can see, the table above is the look we're aiming for after preprocessing the data from the "Titanic - Machine Learning for Disaster Prediction" dataset. Here we have used all preprocessing steps like data description, missing values handling, feature expansion. We were able to verify that errors and fluctuations in the original data set were reduced and numbered in the preprocessed dataset. The preprocessed dataset will be used for training a machine learning model with accuracy and consistency of preprocessed data.

V. CONCLUSION

Through this proposed method, we learned to understand a “clean” and “clean” database that is ready to use in statistical analysis. We have included data cleaning, data integration, and data reduction steps to ensure accuracy. Basic techniques can be applied to solve common problems with raw data, including missing data and data from multiple sources. Data preprocessing step is necessary to solve noisy data, redundant data, missing data values, etc. All learning algorithms rely primarily on the result of this process, which is the final training set and we can do so by selecting relevant cases and working in the same. High-quality data will lead to high-quality results and lower data mining costs. When the data set is too big, the machine algorithm may not be able to run. In this case, instance selection reduces the data and allows the machine learning algorithm to run and work efficiently with big data. In most cases, the missing data must be preprocessed to allow the dataset to be processed by the supervised machine learning algorithm.

Reference

[1] https://medium.com/analytics-vidhya/data-visualization-titanic-data-set -91531c3ab5a6
[2] https://www.researchgate.net/publication/228084519_Data_Preprocessi ng_for_Supervised_Learning.
[3] C. Cardie. Using decision trees to improve cased-based learning. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1995.
[4] Hernandez, M.A.; Stolfo, S.J.: Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data Mining and Knowledge Discovery 2(1):9-37, 1998.
[5] Friedman, J.H. 1997. Data mining and statistics: What’s the connection? Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics.

[6] S. K. Dwivedi and B. Rawat, "A review paper on data preprocessing: A critical phase in web usage mining process," 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 2015, pp. 506-510, doi: 10.1109/ICGCIoT.2015.7380517.
[7] Bauer, K.W., Alsing, S.G., Greene, K.A., 2000. Feature screening using signal-to-noise ratios. Neurocomputing 31, 29–44.
[8] M. Boulle. Khiops: A Statistical Discretization Method of Continuous Attributes. Machine Learning 55:1 (2004) 53-69
[9] Breunig M. M., Kriegel H.-P., Ng R. T., Sander J.: 'LOF: Identifying Density-Based Local Outliers', Proc. ACM SIGMOD Int. Conf. On Management of Data (SIGMOD 2000), Dallas, TX, 2000, pp. 93-104.
[10] Brodley, C.E. and Friedl, M.A. (1999) "Identifying Mislabeled Training Data", AIR, Volume 11, pages 131-167.
[11] https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c 825
[12] https://medium.com/@yogeshojha/data-preprocessing-75485c7188c4
[13] J. Hua, Z. Xiong, J. Lowey, E. Suh, E.R. Dougherty. Optimal number of features as a function of sample size for various classification rules.Bioinformatics 21 (2005) 1509-1515
[14] Isabelle Guyon, André Elisseeff; An Introduction to Variable and Feature Selection, JMLR Special Issue on Variable and Feature Selection, 3(Mar):1157--1182, 2003.
[15] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. Proc. of the 8th International Conference on Machine Learning, 2001.