International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET229626

Framework For Data Development in Mordern Situation Using Machine Learning Technology

Syed Aamir Bokhari, Progyajyoti Mukherjee, Suhail Shaik, Abhishek Samar Singh
Department of Artificial Intelligence and Machine Learning, New Horizon College of Engineering, Bangalore, India

Article Info
Volume 9, Issue 6
Page Number : 229-234

Publication Issue :
November-December-2022

Article History
Accepted : 10 Nov 2022
Published: 28 Nov 2022

ABSTRACT

This research is mainly focused on to talk about data preparation, what better way to start than from observation. Everyone is familiar with the adage that a data scientist should spend 80% of his or her time preparing the data and just 20% actually working with it, particularly when it comes to visualization. This essay will concentrate on data preparation, including the most common issues, solutions, and developments. Data must be put into the proper form before analysis can be done on it. Data manipulation and organization are steps in the preparation of data for analysis. Iteratively transforming unstructured, chaotic raw data into a more organized, practical form that is ready for further analysis is known as data preparation. Data profiling, cleaning, integration, and transformation are just a few of the primary activities (or tasks) that make up the entire preparation process.

Keywords : Data Profiling, Cleaning, Integration, Transformation

I. INTRODUCTION

You probably noticed that data is present everywhere and is what motivates digital innovation. A number of variables, including the proliferation of applications, the increasing importance of the Internet in our daily lives, and the emergence of the IoT (Internet of Things), explain why certain activities are revolving around data. As a result, new positions such as data scientist, data engineer, and data visualization specialist have emerged in IT departments. They are all involved in various ways in the data improvement process. However, they all share the necessity for high-quality data. That is the main goal of data preparation. The process of transforming raw data into something that can be analyzed is referred to as "data preparation.". Importing the data, ensuring its consistency, fixing quality issues, and, if necessary, enriching it with other datasets are all steps in the often tedious process of data preparation. Each stage is crucial and calls for certain functionality, particularly when it comes to data transformation. Data preparation must be carried out within acceptable bounds. Perfectionism, in the words of Winston Churchill, is the adversary of advancement. Without getting bogged down in analysis paralysis or spending infinite hours trying to produce perfect data, the objective is to make the data suitable for its intended use[1]. It cannot, however, be disregarded or left to chance. Understanding the difficulties that data preparation offers and how to deal with them are crucial for success. Although many data preparation

concerns could be grouped under the umbrella term "data quality," it is helpful to separate them into more specialized difficulties in order to find, address, and manage the issues.

II. METHODOLOGIES

Although each data preparation strategy should be tailored to the firm it is intended for, below is a brief description of some typical data preparation procedures. We may divide data preparation into four crucial steps:

1. Learn your Data
2. Validate and Cleanse the Data
3. Improve Data
4. Share Data

1. Learn Your Data

Knowing what you have will help you enhance your data preparation procedures. With spending on data discovery tools expected to increase 2.5 times faster than that of conventional IT solutions, data discovery has emerged as a key priority for investment. 'Discovering' your data simply refers to getting to know it better[2]. Relevant inquiries can include "how am I gathering it" and "what do I want to learn from my data." Successful data analysis depends on having the right data collection strategy.

2. Validate and Cleanse the Data

In essence, this is what we have been discussing throughout the article. Cleaning your data and addressing any mistakes is typically the most time-consuming phase in any data preparation procedure. Standardizing the data entails ensuring that its format is clear, deleting unused or redundant entries, and adding any missing values. Helpful data preparation tools can be very useful in this situation since they can identify inefficiencies and repair faulty formatting.

3. Improve Data

Your method of data preparation is very important at this point. You may now enrich (meaning, improve) your data by adding whatever is missing based on the now-better-defined objectives you came up with in the discovery process. Let's say you want to learn more about any functionality issues your clients may be experiencing. For instance, how effectively the vacuum's batteries serves customers. Combining customer assistance data with customer review data will enhance it; in particular, you should keep track of any reviews that reference the battery[3]. You now have a complete picture of how the battery affects the level of consumer satisfaction with your vacuum.

4. Share Data

It's time to keep your clean, useful data once you've prepared it. We advise choosing a cloud-based storage strategy that is future-proof so you can always modify the data preparation settings for later analysis. In keeping with the theme of being future-ready, let's finish with a list of well-known data preparation tools that may support any data preparation strategy.

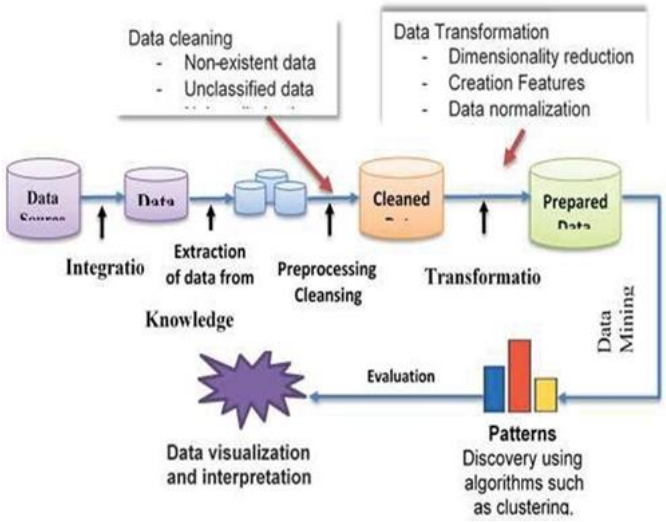

Figure 1 . Data Preparation Structure

Figure 1 depicts the procedure for preparing the data. The initial data analysis plan, which was developed during the research design phase, serves as the overall process's compass. The first step is to look for

questionnaires that are appropriate. The data is then edited, coded, and transcripted. The data are cleaned, and a remedy is suggested for the absence of responses. After the sample validation step, it's common for the data to need statistical adjustment in order to make them representative of the target population. The next step is for the researcher to decide on a suitable data analysis plan. Due to the information and insights learned since the preliminary plan was created, the final data analysis strategy differs from the preliminary plan of data analysis.

III. ALGORITHM

Data cleaning aims to produce straightforward, comprehensive, and understandable sets of examples for machine learning. Particularly in the healthcare industry, real-world problem data is rarely accurate and thorough. Missing values, mistakes, and inconsistencies in the dataset are dealt with during the data cleansing/cleaning stage. Since the past, numerous ways or procedures have been developed to address this issue. Various considerations determine which technique is most appropriate.

The term "missing value" (MV) refers to a data value that is missing from a cell in a certain column. In the context of healthcare, missing values can occur for a variety of causes, some of which are listed below: missing value as a result of human error, not applicable, patient condition unrelated to a particular variable, not electronically captured by the sensor, patient not present on the ventilator as a result of a medical decision, due to an electrical failure, database synchronization, etc.[5]. Working with missing values might provide biased or undesirable results, which can lead to incorrect conclusions. Throwing Out the Missing Values The most common method is to throw away the MVs, although this method is not very useful because if the train data have a lot of missing values, the output must be skewed. If there are only a few missing values in the dataset, we must ensure that the analysis of the remaining data will not result in an inference bias. The MVs can be deleted in the following ways:

Deleting Rows: In this scenario, all detected cases with more than one missing value are eliminated following a thorough case analysis. However, this method only works in a limited percentage of missing person cases. When a dataset contains the MCAR missing pattern, which is uncommon , it functions well.

Pairwise deletion: This technique aims to reduce list-wise deletion's mistake. When examining the data, the attribute containing MVs is destroyed if it is not used as a case for another attribute. Although it increases analysis power, it also introduces other challenges, such as producing standard error.

Completely dropping an attribute: This is really uncommon, however in my opinion there are situations when you can do so if there are more than 60% of observations missing and the property appears to be unimportant in the analysis. Due to their great relevance, some properties with missing values should occasionally be maintained.

Data Transformation: The data is changed into a format at this stage. Within a specific range, the independent data is normalized. This level of modification aids in reducing computing time and accelerating calculations[4]. Data transformations can be done in many different ways, some of which are described below:

Data scaling: This technique returns values between 0 and 1. It is applied to data normalization. For this, the MinMaxScaler class function in Python is employed. this method is used in a variety of classification and regression algorithms, including KNN, neural networks, and others.

Data Standardization: This procedure aids in formatting the data by locating the z-score for input datasets. The input data set is standardized to the range of characteristics in this process.

The result of standardization is that all features will have the same scale, a mean equal to zero, and a standard deviation equal to one. The StandardScaler class in Python has a function called standardization.

Data normalization is a technique used when the distribution of the data is uncertain, when it is non-Gaussian (a bell curve), or when the scale of the data varies. It does not make any assumptions about the distribution of the data and instead rescales the observations to a length of 1. To conduct normalization in Python, use the Normalizer()class. This function can convert inputs to unit norms that can be used for categorization. It is acquired from the Sklearn library[6].

One Hot Encoding: Working with categorical data in machine learning is particularly challenging. It is necessary to transform these classified data into numerical data. We employ the "One Hot Encoding function" to do this. Binary vectors are used in this function to represent categorical variables[8]. Here, categorical values are converted to integer values at first. Each integer value is represented as a binary vector, with the exception of the integer index, where all zero values are designated as 1. The categories are derived using the one-of-a-kind values in each feature by the OneHotEncoder() function, which is defined in the Python Sklerarn package.

Label Encoding: Using this technology, the training data is provided with a text-based label to make it readable. Then, for the machine-readable version, these word labels are transformed into numbers, with the categorical value being replaced by a number between 0 and the number of classes minus 1. On these labels, the machine learning algorithms will operate. In supervised learning, label encoding is a crucial preprocessing step for the structured dataset. The scikit-learn library's LabelEncoder() method is used to conduct label encoding in python.

Smoothing: Using certain algorithms, this approach removes noise from the dataset. Significant features in the dataset are highlighted and patterns are predicted by smoothing. Any variation or other noise source is eliminated or reduced with this technique.

The outliers are eliminated via robust data scaling. By removing the median value, this methodology scales the data according to the interquartile range (IQR). The features that are resistant to outliers are screened by this function. The RobustScaler()class in the scikit-learn Python machine learning toolkit provides access to the robust scaler transforms.

Data reduction: One of the primary criteria in data analysis is the dimension of the data. Data reduction is the process of reducing the dimension of the data in order to boost data efficiency and make storage of the data more manageable. This method has been applied to different data analysis fields as well as to escape the dimensionality curse. Dimension reduction (DR) is a strategy that is inversely correlated with the dimensionality curse. The "curse of dimensionality" is mostly a result of insoluble issues brought on by lengthy computations required for data processing[7]. It has been found that the cost of computation exponentially rises with the number of variables. DR techniques have thus been employed to avoid this issue. Popular DR techniques include factor analysis and principal component analysis (PCA) (FA).

In order to get around the "curse of dimensionality," these strategies decrease the amount of variables in the dataset and progressively lengthen the computation time.

In order to reduce the number of correlated variables (p) to fewer k (k p) uncorrelated variables, the PCA technique conducts a linear dimensionality reduction. Principal components are these uncorrelated variables. The primary purpose of PCA is to identify features that are associated with one another. If the correlation value is more than a threshold, the function will combine the features and create data for the remaining, linearly uncorrelated features. By identifying the largest variance in the initial high-dimensional data and projecting them onto a less dimensional space, the PCA method operates until correlation is somewhat reduced.

Feasibility Analysis (FA) Similar to PCA, factor analysis is used to uncover hidden variables in datasets—variables that are not directly assessed in one variable but are instead produced by other variables—and to reduce the dimensionality of the data. Factors are the name given to these hidden variables[9].

ML runs on data. While difficult, using this data to reinvent your company is essential for both the now and the future. Survival of the most knowledgeable applies, and those who can use their data to make better, more educated judgments are more likely to be able to react quickly to the unexpected and find new opportunities. This crucial yet time-consuming procedure is a requirement for creating accurate ML models and analytics, and it takes up the majority of an ML project's time. Data scientists can utilize a variety of tools to help automate data preparation in order to reduce the time commitment.

IV. CONCLUSION AND FUTURE ENHANCMENT

Data analysts use data preparation today to guide their quality knowledge discovery and help create efficient, high-performance data analysis application systems. In data mining, the data preparation stage is in charge of selecting high-quality data from the pre-processed data. Because real-world data is imperfect, high-performance mining systems demand quality data, and quality data produces concentrative patterns, data preparation is crucial.

In this essay, we have argued for the significance of data preparation and provided a quick overview of the field's research; the specifics of each accomplishment may be found in this special issue. In order to summarize, we will now talk about potential directions for data preparation. There are a lot of difficult research questions for data preparation due to the variety of data and data-mining jobs. The following are some potential directions for data preparation in the future:

1) Building environments for interactive and integrated data mining.
2) Stablishing theories for data preprocessing.
3) Creating methods and algorithms for data preparation that are efficient and effective for both single and many data sources while taking into account internal and external data
3) Investigating effective methods of data preparation for Web intelligence

A research that indicated that 74% of data scientists stated it is the toughest aspect of their professions is proof that data preparation has a negative image. In the future, we must develop a method or algorithm that can swiftly and efficiently clean the data and save a ton of time.

V. REFERENCES

[1] Zhang, Z., C. Zhang, and S. Zhang. 2003. An agent-based hybrid framework for database mining. Applied Artificial Intelligence 17(5–6):383–398.
[2] Zhang, C., and S. Zhang. 2002. Association Rules Mining: Models and Algorithms. In Lecture Notes in Artificial Intelligence, volume 2307, page 243, Springer-Verlag

[3] Zhang, H., and C. Ling. 2003. Numeric mapping and learnability of Na€ve Bayes. Applied Artificial Intelligence 17(5–6):507–518
[4] Yang, Q., T. Li, and K. Wang. 2003. Web-log cleaning for constructing sequential classifiers. Applied Artificial Intelligence 17(5–6):431–441.
[5] Tseng, S., K. Wang, and C. Lee. 2003. A pre-processing method to deal with missing values by integrating clustering and regression techniques. Applied Artificial Intelligence 17(5–6):535–544
[6] Ratanamahatana, C., and D. Gunopulos. 2003. Feature selection for the Naive Bayesian classifier using decision trees. Applied Artificial Intelligence 17(5–6):475–487
[7] Hruschka, E., Jr., E. Hruschka, and N. Ebecken. 2003. A feature selection Bayesian approach for extracting classification rules with a clustering genetic algorithm. Applied Artificial Intelligence 17(5–6):489–506.
[8] Batista, G., and M. Monard. 2003. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence 17(5–6):519–533.
[9] Abdullah, N., M. Liquie`re, and S. A. Cerri. 2003. GAsRule for knowledge discovery. Applied Artificial Intelligence 17(5–6):399–417.

Cite this article as :

Syed Aamir Bokhari, Progyajyoti Mukherjee, Suhail Shaik, Abhishek Samar Singh, "Framework For Data Development in Mordern Situation Using Machine Learning Technology", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 6, pp. 229-234, November-December 2022. Available at doi : https://doi.org/10.32628/IJSRSET229626
Journal URL : https://ijsrset.com/IJSRSET229626