International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 231-235, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1008

Data Preparation for Machine Learning Modelling

Ndung’u Rachael Njeri
Information Technology Department
Murang’a University of Technology
Murang’a, Kenya

Abstract: The world today is on revolution 4.0 which is data-driven. The majority of organizations and systems are using data to solve problems through use of digitized systems. Data lets intelligent systems and their applications learn and adapt to mined insights without been programmed. Data mining and analysis requires smart tools, techniques and methods with capability of extracting useful patterns, trends and knowledge, which can be used as business intelligence by organizations as they map their strategic plans. Predictive intelligent systems can be very useful in various fields as solutions to many existential issues. Accurate output from such predictive intelligent systems can only be ascertained by having well prepared data that suits the predictive machine learning function. Machine learning models learns from data input using the ‘garbage-in-garbage-out’ concept. Cleaned, pre-processed and consistent data would produce accurate output as compared to inconsistent, noisy and erroneous data.

Keywords: Data Preparation; Data pre-processing; Machine Learning; Predictive models

1. INTRODUCTION

The world is witnessing a fourth industrial revolution, which is fast-paced due to technological evolutions and advancements. Today, digital systems are been experienced in all spheres of the industries including and not limited to healthcare, education, manufacturing, entertainment, and telecommunication where there’s a wealth of data. The digital systems have become sources of massive data, where insights can be extracted and analyzed for new patterns and new knowledge that may be useful in building various smart applications in the pertinent domains.

2. Data Pre-processing

Data pre-processing is an important step while developing smart systems or while extracting meaningful insights using machine learning. Data processing is sometimes used interchangeably with data preparation; however, data processing is inclusive of both data preparation and feature engineering whereas data preparation excludes feature engineering [4]. Before data preparation, there is usually need to understand the output you require from the machine model to be trained, and hence the subsequent data attributes that will shape the output. With the output in mind, the data to be collected is easily identifiable, and thus its quality and value requirements defined. This problem articulation ascertains the right steps of data preparation are followed.

The data pre-processing involves data cleaning, which involves removal of ‘dirt’ or noise in data, removal of missing or inconsistent data, data integration if data is sourced from multiple sources, data transformations depending on the type of raw data to what the machine learning algorithms can use as inputs, data reduction where unnecessary data is removed and only data that is required to develop an application is retained [5]. Data pre-processing makes sure that the data types to use in machine learning functions are transformed, an imposition requirement by some machine learning algorithms on data, with some having non-linear relationships that complicates how the algorithms functions [6].

2.1 DATA PREPARATION

Data preparation is the process of converting raw data through pre-processing before being used in fitting and evaluating machine learning predictive systems [6]. Machine learning models are particular to their data source, and hence the credibility of the data source and utility of the data collected is essential. It is plausible for a machine learning model to be high end model but training it with the wrong data yields the wrong information. Machine learning models operate on the “garbage in, garbage out” philosophy, and data scientists ensure the “garbage in” remains relevant, for the resultant information to be relevant. Standardizing your data entry point ensures the right information is attained at the end result. For these reasons, data collection remains an imperative part of data preparation.

Data preparation ascertains minimal errors in your data, and allows for data monitoring of any future errors. This will eventual ensure the machine learning is trained with the correct data and hence the output will be accurate. Data exploration analysis will provide a summary of your data set, and allow for necessary changes or formatting to be done. Any data source in machine learning is divide into both the training and the test data, and the technique of this division is achieved during data preparation. Additionally, data preparation helps in shaping the data to fit the requirements of the machine learning model.

Some data sets have attributes that are not well ordered for analysis. Other times, the ranges in the data sets to be compared largely vary, resulting to comparison challenges. Data transformation allows for such data sets to be transformed into good representations of the initial data source, without losing data relevancy or data integrity. Some training models accept input data in certain formats, necessitating data transformation.

In an era of big data, there is need to create better storage techniques and often times this is costly, both in terms of storing the big data, and in analyzing it. Big data analytics require complex software which is expensive. Data reduction comes in handy in compressing data into more manageable volumes while retaining its relevance and integrity. Additionally, the reduced volumes can be used in computations as a representation of the whole data set with trivial to zero impact on the initial data source, and the output of the model. Data reduction reduces the overall cost of data analysis, and saves on the time that would have otherwise been employed in future data processing.

The main four steps for data preparation are data collection, data cleaning, data transformation and data reduction.

2.2 DATA COLLECTION

Data collection is the initial stage of data preparation, and it involves deciding on the data set depending on the expected output of the machine model to be trained. Essentially, collection of the right data set ascertains the right data output. Data collection consists of data acquisition, data labeling, data augmentation, data integration and data aggregation.

2.2.1 Data acquisition.

Data acquisition involves identifying the data source, defining the methodology of collecting the data, and converting the collected data into digital form for computation. The data source can be primary, where data is obtained straight from the persons, objects or processes being studied. When your data In this stage, exploratory data analysis (EDA) is used, and it is a technique that aims at understanding the characteristics and attributes of the data sets [12]. It aids in the data scientist becoming more familiarized with the data collected. In exploratory data analysis, statistical tools and techniques are applied in building hypothesis source is a party that had previously collected data, it is termed as a secondary source. Methodology of data collection varies depending on the expected output. Statistical tools and techniques are applied in both the collection of qualitative and quantitative data.

2.2.2 Data labelling

As machine learning advances, there is development of deep learning techniques which have automated the generation of features from data sets, and hence the requirement of high volumes labelled data [7]. Data labelling is the process through which the data models are trained through tagging of data samples. For instance, if a model is expected to tell the difference between images of cats and dogs, it will be initially introduced to images of cats and dogs, which are tagged as either cats or dogs. This is done manually, though often with the aid of a software. This part of supervised learning allows the model to form a basis of future learning. The initial formation of a pattern in both the input and output data, defines the requirements of the data to be collected. Therefore, before data collection is initialized, there is need to delineate the data parameters and the intended information to be retrieved from the data.

2.2.3 Data augmentation

Data augmentation is a data preparation strategy that is used in increasing data diversity for deep learning model training [8]. It involves construction of iterative optimization with the aim of developing new training data from already existing data. It allows for the introduction of unobserved data or introduction of variables that are inferred through mathematical models [9]. While not always necessary, it is essential when the data being trained is complex and the available volume of sampled data is small. Data augmentation saves the problem of limited data and model overfitting [10].

2.2.4 Data aggregation

Data aggregation is a technique of reducing the volume of data though grouping. This grouping is usually of a single attribute. For instance, when one has a data set with the attribute time organized in days over a given time series, one can aggregate the data into monthly groups which eases dealing with the time attribute. It aids in reducing the broadness of a given attribute without tangible losses during future data manipulation [10].

2.3 DATA CLEANING

Data cleaning, also referred to as data cleansing is the technique of detecting and correcting errors and inaccuracies in the collected data [11]. Data is supposed to be consistent with the input requirement of the machine learning model. The main activities in data cleansing involve the fine-tuning of the noisy data and dealing with missing data. It aids in ensuring the collected data set is comprehensive and any errors and biases that may have arose in data collection have been eliminated. This includes the detection of outliers within the data set; both for the numerical and the non-numerical data sets.

2.3.1 Exploratory Data Analysis

on the information that can be attained from the collected data, and sometimes involves data visualization. Data visualization allows for the understanding of data properties as skewness and outliers.

Exploratory data analysis is mainly done on the statistical manipulation software. The graphical techniques allow for understanding the distribution of the data set, and the statistical summary of all attributes. EDA allows for future decisions such as the data cleansing techniques to be used, what data transformations are necessary and whether data reduction is necessary and if yes, what is technique to use. Exploratory data analysis is a continuous process all through data preparation.

2.3.2 Missing Data

While it is important to ascertain during data collection that all the attributes of the data sets have their real value collected, data sometimes has some of the attributes with missing values, which makes it hard to use as input in machine learning models. As so, different techniques have been outlined on how to deal with missing data. Data manipulation platforms as python and R statistics have some of these techniques of dealing with missing data embedded in them. The best technique usually varies with the data set, and hence after data assessment in the exploratory data analysis, one can easily select the best technique for missing data imputation.

2.3.2.1 Deductive Imputation

Deductive imputation follows the basic rule of logic, and is hence the easiest imputation, however, the most time consuming. Even so, its results are usually highly accurate. For instance, if student data indicates that the total number of students is 10, and the total number of examinations papers is 10, but there is a paper with a missing name and John has no marks recorded, logic dictates the nameless paper is John’s. However, deductive imputation is not applicable in all types of data sets [13].

2.3.2.2 Mean/Median/Mode Imputation

This imputation uses statistical techniques where the central measures of tendency within a certain attribute are computed and the missing values replaced with the computed measure of central tendency, may it be mean, mode or the median of that attribute [13]. This technique is applied in numerical data sets, and the impact on the output or later computations is trivial. Data manipulation platforms as python and R statistics have techniques of dealing with missing data embedded in them.

2.3.3 Noisy Data.

Presence of noisy data can have substantial effect on the output of a machine model. It negatively impacts on prediction of information, ranking results, and the accuracy in clustering and classification [14]. Noisy data includes unnecessary information in the data, redundant data values and duplicates or pointless data values. These result from faultiness in collection of data, problems that may result from data entry, problems that occur from data transfer techniques applied, uneven naming conventions of the data and sometimes it may arise from technology restrictions, as in the case of unstructured data. Noisy data is eliminated through.

2.3.3.1 Binning Method

This involves arranging data into groups of given intervals, and is used in smoothening ordered data. The binning method relies on the measures of central tendency and it is done in one of three ways. Smoothing by bin means, smoothing by bin median and smoothing by bin boundary.

2.3.3.2 Regression

Linear Regression is a statistical and supervised machine learning technique, that predicts particular data based on existing data [15]. Simple linear regression is used to compute the best line of fit based on existing data, and hence outliers in the data can be identified. To attain the best line fit, there is development of the regression function based on the prior collected data. However, it is important to note that though in some data sets, extreme outliers are considered noisy data, the outliers can be essential to the model.

For instance, if an online retailer company has its market within countries in Europe and trivial market in the United States, the United States may be considered an extreme outlier, and hence noisy data. However, a machine learning model may realize that though a very small number of the Americans use the online platform, they bring in more revenue than some of the countries in Europe. Simple linear regression uses one independent variable whereas multiple linear regression uses more than one independent variable in its computations.

2.3.3.3 Clustering

Clustering is in the unsupervised machine learning category and it operates by basically grouping the collected data set into clusters, based on their attributes (Gupta & Merchant, 2016). In clustering, the outliers in the data may fall within the clusters, and in the case that they are extreme outliers they fall outside the clusters. To understand the effect of clustering, data visualization techniques are used “Clustering methods don’t use output information for training, but instead let the algorithm define the output” [17]. There are different techniques used in clustering.

In K-means clustering, K is the number of clusters to be made, and to do this the algorithm randomly selects K number of data points from the data set. These K data points are called the centroids of the data, and every other data point in the data set is assigned to the closest centroid. This process is repeated for all the new K data sets created, and the process iterated until the centroids become constant, or fairly constant. This is called the point at which convergence occurs. The Density-Based Clustering of Applications with Noise (DBSCAN) is used in data set smoothing.

2.4 DATA TRANSFORMATION

Data transformation involves shifting the cleansed data from one format to the next, from one structure to the next, or changing the values in the cleansed data set to meet the requirements of the machine learning model [18]. The simplicity of the data transformation is highly dependent on the required data for input, and the available data set. Data transformation involves:

2.4.1 Normalization

Normalization is a technique for data transformation that is applied in numeric values of columns when there is for a common scale. This transformation is achieved without loss of information, but only changing how it is represented. For instance, in a data set with two columns that have different scales such as one with values ranging from 100 to 1,000 and another column with a value range of 10,0000 to 1,000,000 there may arise a difficulty in the event that the two columns have to be used together in machine learning modelling. Normalization finds a solution by finding a way of representing the same information without loss of distribution or ratios from the initial data set [19].

It is imperative to note that while normalization is only necessitated by the nature of some data sets, other times it is demanded by the machine learning algorithms being used. Normalization uses different mathematical techniques such as z-score in data standardization. The technique picked is usually decided depending on the nature and characteristics of the dataset. Therefore, it is decided at the exploratory data analysis stage.

2.4.2 Attribute selection

In this transformation, latent attributes are created based on the available attributes in the data set to facilitate the data mining process [18]. The latent attributes created usually have no impact on the initial data source, and therefore can be ignored afterwards. Attribute transformation usually facilitates classification, clustering and regression algorithms. Basic attribute transformation involves decomposition of the available attributes through arithmetic or logical operations. For instance, a data set with a time attribute given in months, can have its month attribute decomposed to weeks, or aggregated to years depending on the requirements.

2.4.3 *Discretization*

In data transformation by discretization, there is creation of intervals or labels, and eventual mapping of the all data points to the created data intervals or labels. The data in question is customarily numeric data. There are different statistical techniques used in discretization of data sets. The binning method is used on ordered data, where the data is creation of data intervals called bins where all the data points are mapped into. In data discretization by histogram analysis, histograms are used in dividing the values of the attribute into disjoint ranges where all other data points are mapped to. Both binning and histogram analysis are unsupervised data discretization methods.

In data discretization by decision tree analysis, the algorithm picks the attribute with the minimum entropy, and uses its minimum value as the point from which it, in iterations, partitions the resulting intervals till it attains as many different groups as possible [20]. This discretization is hierarchical hence its name. To use an analogy, it’s like dividing a room into two equal parts, and continuously dividing the resulting partitions into two other equal parts. Only in this case, the room has multi-varied contents and we want each different content in its own space at the end of the partitioning. This discretization technique uses a top-down approach and is a supervised algorithm.

Data discretization by correlation analysis is highly dependent on mathematical tools and it applies a bottom-up approach, unlike decision trees [20]. It maps data points to data intervals by the best neighboring interval for each data point, and merging the intervals. It then recursively repeats the process to create one large interval. It is a supervised machine learning methodology.

2.4.4 Concept Hierarchy Generation

In concept hierarchy data transformation, there is mapping of low-level concepts within the attributes to higher level concepts [21]. Most of these concepts are normally implied in the initial data set, and hence the technique is embedded in statistical software. It follows a bottom up approach. For instance, in the location dimension, cities can be mapped to their states, their provinces, their countries and eventually their continents.

.

2.5 DATA REDUCTION

With the advancement of trends in information technology and the exponential growth of internet of things, there has been an eventual precipitous increase in the volumes of available data. This is a huge benefit to machine learning as the availability of big data for training the models ascertains accuracies in the outputted information from such models. Nonetheless, handling and analyzing these enormous volumes of data is a big challenge, hence the need for data reduction techniques. Data reduction reduces the cost of analyzing and storing these volumes of data by increasing storage efficiency. The different techniques used in data reduction include.

2.5.1 Data cube aggregation

A data cube is an n-dimensional array that uses mathematical tensors to represent information. the online analytical processing (OLAP) cube stores data in a multidimensional form, which occupies lesser storage space compared to a unidimensional storage technique [22]. To access data from the OLAP cube, the Multidimensional expressional (MDX) query language is used. The query language includes the roll-up, drill-down, slice and dice and pivot operations. These operations allow access to the required attributes of the data from the cube, without removing the data from the data cube, hence saving on space.

2.5.2 Attribute subset selection

Attribute subset selection, also known as feature selection is a part of feature engineering and it involves the discovery of the smallest possible subset of attributes that would yield the same results or closest to the same results on data mining, as when using all the attributes [23]. This technique ensures that only what is completely necessary from the initial data set is used in the modeling. This simplifies detection of insights, patterns and information from the data set while saving on analysis and storage costs.

2.5.3 Numerosity reduction

In numerosity reduction data reduced and made feasible for analysis through replacement of the original data with a model of the data that preserves the integrity of the initial data [24]. Two statistical method are used in the creation of the representational model. In the parametric method, regression and log-linear methods are sued in the development of the representational model. Non-parametric methods encompass the use of clustering, sampling, use of histograms and data cube aggregation to represent the whole data population, during computations and storage.

3. POSSIBLE BIASES IN DATA PREPERATION

Bias in the data to be trained in the machine learning model leads to consequent wrong information output. It is imperative to identify the source of any bias in your data set during data preparation and eliminate the bias [25]. Sample bias occurs at data collection where the selected data sample is not the right representation of the population under study, hence it is also called selection bias. For instance, an iris scan recognition trained entirely on the iris scans of Africans will not efficiently identify eyes of the white population.

Exclusion bias is common in the data cleansing stage where there is deletion, or misrepresentation of a part of the data, leading to it being excluded in the model training. Measurement bias occurs either during data collection, where the system of collecting input data is not the same as that of collecting output data. Additionally, it occurs during data labelling, where non-uniform data labelling results to faulty predictions from the machine learning model. Recall bias also occurs at the data labelling stage, where the labelling is non-consistent [25].

Observer bias is data fallacy where the person dealing with the data assumes the observation to be wat they expected, as opposed to the real observation. Data scientists and researchers are encouraged to operate on an objective rather than subjective approach to avoid this bias [19]. Another is racial bias, and the best example of this bias in talk balk engines, where the model was largely trained on the voice data of the white population, and hence it hardly recognizes the voice of the black data population [19]. Association bias occurs when a data set has created an implicit association between attributes. The main association bias is the gender bias, as in the case where a system is trained with all school principals being males, and hence eventually disqualifies the plausibility of a female school principle [25].

4. CONCLUSION

Many machine learning predictive systems and models are affected by the kind of data that is used as input of the models. Results of the predictive models are determined by the machine learning algorithm function and the kind of data input. Biased data will produce biased results. Equally, ‘dirty’ data will produce wrong results or output that cannot be relied upon.

It’s imperative to have clean data to fit in the machine learning models so as to have the models learn correctly and predict accurately. There is high chance that inaccurate results from machine learning models are caused by improperly prepared input data. Therefore, for ensuring the explainability and reliability of machine learning predictive models that are used to develop intelligent systems, clean prepared data is significant.

Digital data sources such as internet of things which is a major source of real-world data have noisy, inconsistent and missing data, which when used in predictive modelling using machine learning functions can result to erroneous and inaccurate results. Removal of such inconsistencies in input data cannot be overemphasized. Clean data which is formatted and organized to the required standard of the machine learning function goes a long way in contributing towards better machine learning models with reliable results. There is more to data preparation than has been included on this work. In future, we look to define different types of data and their various pre-processing methods.

5. ACKNOWLEDGMENTS

My thanks to all authors whom I have referenced here below for their research works, which was insightful and helped to compile above findings.

6. REFERENCES

[1] applications and research directions." *SN Computer Science* 2, no. 3 (2021): 1-21.

[2] Altexsoft. (2018, June 16). Preparing Your Dataset for Machine Learning: 8 Basic Techniques That Make Your Data Better. Retrieved on July 29, 2020 from: *https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/*

[3] Bengfort, B., & Kim, J. (2016). Data analytics with Hadoop: an introduction for data scientists. " O'Reilly Media, Inc.".

[4] El-Amir, H., & Hamdy, M. (2020). Data Wrangling and Preprocessing. In Deep Learning Pipeline (pp. 147-206). Apress, Berkeley, CA. Retrieved from: *https://doi.org/10.1007/978-1-4842-5349-6_*

[5] García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72, pp. 59-139). Cham, Switzerland: Springer International Publishing.

[6] Brownlee, J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

[7] Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*. Retrieved from: *https://ieeexplore.ieee.org/abstract/document/8862913*

[8] Ho, D., Liang, E., Liaw. R., (2019, June 7). 1000x Faster Data Augmentation. Berkeley Artificial Intelligence Research. Retrieved on July 29, 2020.

[9] Antoniou, A., Storkey, A., & Edwards, H. (2017). Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340.

[10] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60. Retrieved from: https://doi.org/10.1186/s40537-019-0197-0

[11] Murata, K., Noda, H., & Haraguchi, M. (2017). U.S. Patent No. 9,558,151. Washington, DC: U.S. Patent and Trademark Office. Retrieved from: *https://patents.google.com/patent/US9558151B2/en*

[12] Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*(2), 265-276. Retrieved from: *https://doi.org/10.1016/j.hrmr.2016.08.003*

[13] Van der Loo, M., & de Jonge, E. (2017). deductive: Data Correction and Imputation Using Deductive Methods. R package version 0.1, 2.

[14] Gupta, S., & Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science*, *161*, 466-474. Retrieved from: *https://doi.org/10.1016/j.procs.2019.11.146*

[15] Elgabry, O. (2019, March 1st). The Ultimate Guide to Data Cleaning. Retrieved on July 27, 2020 from: *https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4*

[16] Gupta, A., & Merchant, P. S. (2016). Automated lane detection by k-means clustering: a machine learning approach. *Electronic Imaging*, *2016*(14), 1-6. Retrieved from: *https://doi.org/10.2352/ISSN.2470-1173.2016.14.IPMVA-386*

[17] Castañón, J. (2019, May 2nd). 10 Machine Learning Methods that Every Data Scientist Should Know. Retrieved on 26th July 2020, from*: https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9*

[18] Malik, K. R., Ahmad, T., Farhan, M., Aslam, M., Jabbar, S., Khalid, S., & Kim, M. (2016). Big-data: transformation from heterogeneous data to semantically-enriched simplified data. *Multimedia Tools and Applications*, *75*(20), 12727-12747. Retrieved from: *https://doi.org/10.1007/s11042-015-2918-5*

[19] Microsoft. (2020, April, 7th). Bias in Machine Learning. Retrieved on July 31, 2020 from: *https://devblogs.microsoft.com/premier-developer/bias-in-machine-learning/*

[20] Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., ... & Herrera, F. (2016). Data discretization: taxonomy and big data challenge. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *6*(1), 5-21. Retrieved from: *https://doi.org/10.1002/widm.1173*

[21] Swamy, M. K., & Reddy, P. K. (2020). A model of concept hierarchy-based diverse patterns with applications to recommender system. *International Journal of Data Science and Analytics*, 1-15. Retrieved from: *https://doi.org/10.1007/s41060-019-00203-2*

[22] Shen, H., Zhang, M., & Shen, J. (2017). Efficient privacy preserving cube-data aggregation scheme for smart grids. IEEE Transactions on Information Forensics and Security, 12(6), 1369-1381. Retrieved from: https://ieeexplore.ieee.org/document/7828093

[23] Demisse, G. B., Tadesse, T., & Bayissa, Y. (2017). Data Mining Attribute Selection Approach for Drought Modeling: A Case Study for Greater Horn of Africa. *arXiv preprint arXiv:1708.05072*. retrieved from: *https://arxiv.org/ct?url=https%3A%2F%2Fdx.doi.org%2F10.5121%2Fijdkp.2017.7401&v=2a6e454a*

[24] Deepak, J. (n.d.). Numerosity Reduction in Data Mining. Retrieved on July 25, 2020 from: *https://www.geeksforgeeks.org/numerosity-reduction-in-data mining.*

[25] Liam, H. (2020, July 20th). 7 Types of Data Bias in Machine Learning. Retrieved on July 31, 2020 from: *https://lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/*