DATA WRANGLING AND DATA LEAKAGE IN MACHINE LEARNING FOR HEALTHCARE

[1]Saravanan N, [2]Sathish G, [3]Balajee J M
[1 2]Assistant Professor, [3]Research Scholar
[1 2]Department of Computer Application,
[1 2]Priyadarshini Engineering College, Vaniyambadi, Vellore, Tamilnadu, India
[3]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India

Abstract: Nowadays, healthcare and life sciences overall have produced massive amounts of real-time data by enterprise resource planning (ERP). These large of amount data is a difficult task to handle, and intimidation of data leakage by inside worker rises, the firms are smearing way out for security such as Data Loss Prevention (DLP) and Digital Rights Management (DRM) to prevent data leakage. On the other hand, data leakage system also turns into varied and challenging to avert data leakage. Machine learning techniques are used for the handling of significant data by evolving algorithms and set of rules to provide the prerequisite results to the workers. Deep learning has automatic feature extraction that grasps the essential features necessary for the solution of the problem. It reduces the issue of the workers to select elements explicitly to solve the problems for supervised, unsupervised and semi-supervised for healthcare data's.

IndexTerms: Data leakage, Machine Learning, Deep Learning, Healthcare, Enterprise resource planning

Introduction

Deep learning and Machine learning plays a vital role in today's ERP (Enterprise Resource Planning). In the practice of building the analytical model with Deep Learning or Machine Learning the data set is collected as of various sources such as a file, database, sensors and much more [1]. The received data cannot be used openly for carrying out analysis process. To solve this problem Data Preparation is done by using two techniques are data preprocessing and data wrangling [2].

Data Preparation is an essential part of Data Science. It consists of two notions such as Data Cleaning and Feature Engineering. These two are unavoidable for achieving better accuracy and performance in the Machine Learning and Deep Learning tasks [3].

Data Preprocessing is a procedure that is used to transform the raw data into a clean data set. Also, every time the data collected from different sources in raw format which is not viable for the analysis [4]. Therefore, specific phases are executed to convert the data into a small clean dataset. This technique is implemented earlier the execution of Iterative Analysis. The set of steps is wellknown as Data Preprocessing. It includes Data Cleaning, Data Integration, Data Transformation and Data Reduction.

Data Wrangling is a technique executed at the time of making an interactive model. In other words, it is used to convert the raw data into the format that is convenient for the consumption of data. This technique is also known as Data Munging. This method also follows specific steps such as after extracting the data from different data sources, sorting of data using particular algorithm is performed, decompose the data into a separate structured format and finally store the data into another database [5].

In order to achieve better results from the applied model in Machine Learning and Deep Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning and Deep Learning model need data in a specified format, for example, Random Forest algorithm does not support null values, and therefore to execute random forest algorithm null values has to be managed from the original raw data set [6]. Another aspect is that dataset should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one dataset and best out of them is chosen.

Data wrangling is an important aspect for implementing the model. Therefore, data is converted to the proper feasible format before applying any model intro it [7]. By performing filtering, grouping and selecting appropriate data accuracy and performance of the model could be increased.

Another concept is that when time series data has to be handled every algorithm is executed with different aspects. Therefore Data Wrangling is used to convert the time series data into the required format of the applied model [8]. In simple words, the complex data is converted into a usable format for performing analysis into it.

Need of data preprocessing and data wrangling

In order to achieve better results from the applied model in Machine Learning and Deep Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning and Deep Learning model need data in a specified format, for example, Random Forest algorithm does not support null values, and therefore to execute random forest algorithm null values has

to be managed from the original raw data set [9]. An additional phase is that dataset should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one dataset and best out of them is chosen.

Need of Data Wrangling

Data wrangling is an important aspect for implementing the model. Therefore, data is converted to the proper feasible format before applying any model intro it. By performing filtering, grouping and selecting appropriate data accuracy and performance of the model could be increased. Another concept is that when time series data has to be handled every algorithm is executed with different aspects [10]. Therefore Data Wrangling is used to convert the time series data into the required format of the applied model. In another verses, the complex data is converted into a usable format for performing analysis into it.

Use of Data Pre-processing

Data Preprocessing is necessary because of the presence of unformatted real world data. Mostly real world data is composed of

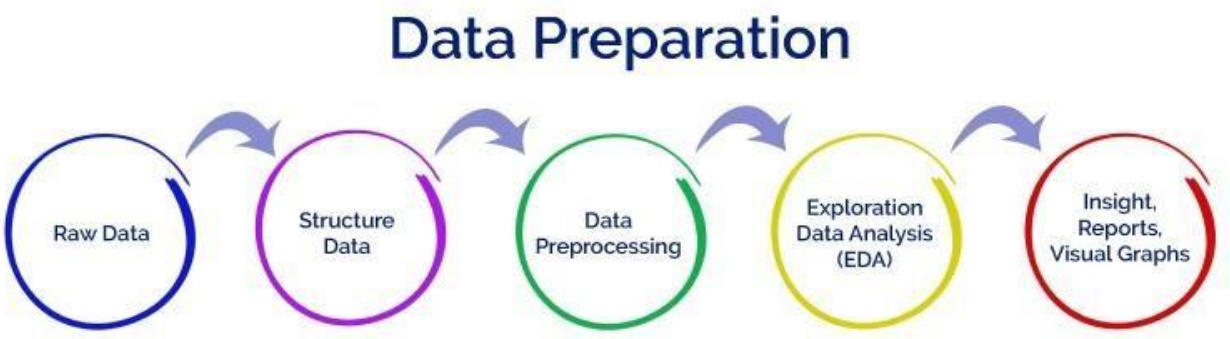

Figure 1. Data Preparation from xeonstack

Inaccurate data (missing data) —There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics and much more.

The presence of noisy data (erroneous data and outliers) —the reasons for the presence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.

Inconsistent data —the presence of inconsistencies is due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names i.e. violation of data constraints and much more.

To handle raw data, Data Preprocessing is performed.

Use of Data Wrangling

Data Wrangling is used to handle the issue of Data Leakage while implementing Machine Learning and Deep Learning.

Data leakage in machine learning / deep learning

Data Leakage is responsible for the cause of invalid Machine Learning/Deep Learning model due to the over optimization of the applied model. Data Leakage is the term used when the data from outside i.e. not part of training dataset is used for the learning process of the model. This additional learning of information by the applied model will disapprove the computed estimated performance of the model [11].

For example when we want to use the specific feature for performing Predictive Analysis but that specific feature is not present at the time of training of dataset then data leakage will be introduced within the model.
Data Leakage can be demonstrated in many ways that are given below:

- Leakage of data from test dataset to training dataset.
- Leakage of computed correct prediction to the training dataset.
- Leakage of future data into the past data.
- Usage of data outside the scope of applied algorithm

The leakage of data is observed from two main sources of Machine Learning/Deep Learning algorithms such as feature attributes (variables) and training dataset [12].

Data Leakage is observed at the time of usage of complex datasets. They are described below:

- At the time of dividing time series dataset into training and test, the dataset is a complex problem.
- Implementation of sampling in a graphical problem is a complex task.
- Storage of analog observations in the form of audios and images in separate files having a defined size and timestamp.

Performance of data pre-processing

Data Preprocessing is performed to remove the cause of unformatted real world data and the missing data to handle [13]. There are three different steps that can be executed which are given below:

- **Ignoring the missing record**—It is the simplest and effective method for handling the missing data. But, this method should not be performed at the time when the number of missing values is huge or when the pattern of data is related to the unrecognized basic root of the cause of statement problem.

Filling the missing values manually—This is one of the best-chosen methods. But there is one limitation that when there is large dataset and missing values are large then, this method is not efficient as it becomes a time-consuming task.

Filling using computed values—The missing values can also be filled by computing mean, mode or median of the observed given values. Another method could be the predictive values that are computed by using any Machine Learning or Deep Learning algorithm. But one drawback of this method is that it can generate bias within the data as the computed values are not accurate with respect to the observed values.

Figure 2. Data Pre-Processing from xeonstack

Process of handling the noisy data

The methods that can be followed are given below:

Binning method—In this method sorting of data is performed with respect to the values of the neighborhood. This method is also known as local smoothing.

- **Clustering method**—In the approach, the outliers may be detected by grouping the similar data in the same group i.e. in the same cluster.

- **Machine Learning**—A Machine Learning algorithm can be executed for smoothing of data. For example, regression algorithm can be used for smoothing of data using a specified linear function.

- **Removing manually**—The noisy data can be removed manually by the human being but it is a time-consuming process so mostly this method is not given priority.

The inconsistent data is managed using external references and knowledge engineering tools like knowledge engineering process.

Data Leakage in Machine Learning

Data leakage can cause you to create overly optimistic if not completely invalid predictive models. Data leakage is when information from outside the training dataset is used to create the model [14]. This additional information can allow the model to learn or know something that it otherwise would not know and in turn invalidate the estimated performance of the mode being constructed.

It is a serious problem for at least 3 reasons:

1. **It is a problem if you are running a machine learning competition**. Top models will use the leaky data rather than be good general model of the underlying problem.
2. **It is a problem when you are a company providing your data**. Reversing an anonymization and obfuscation can result in a privacy breach that you did not expect.
3. **It is a problem when you are developing your own predictive models**. You may be creating overly optimistic models that are practically useless and cannot be used in production.

To overcome there are two good techniques that you can use to minimize data leakage when developing predictive models are as follows:

1. Perform data preparation within your cross validation folds.
2. Hold back a validation dataset for final sanity check of your developed models.

Perform Data Preparation Within Cross Validation Folds

When data preparation of data, the leak of information in machine learning may occur.The effect is over fitting your training data and having an overly optimistic evaluation of you models performance on unseen data. To normalize or standardize your entire dataset, then estimate the performance of your model using cross validation, you have committed the sin of data leakage.

The data rescaling process that you performed had knowledge of the full distribution of data in the training dataset when calculating the scaling factors (like min and max or mean and standard deviation). This knowledge was stamped into the rescaled values and exploited by all algorithms in your cross validation test harness [15].

A non-leaky evaluation of machine learning algorithms in this situation would calculate the parameters for rescaling data within each fold of the cross validation and use those parameters to prepare the data on the held out test fold on each cycle. To re-prepare or re-calculate any required data preparation within your cross validation folds including tasks like feature selection, outlier removal, encoding, feature scaling and projection methods for dimensionality reduction, and more. When performing feature selection on all of the data and then cross-validate, then the test data in each fold of the cross-validation procedure was also used to choose the features and this is what biases the performance analysis.

Hold Back a Validation Dataset

A simpler approach is to split your training dataset into train and validation sets, and store away the validation dataset. After completed your modeling process and actually created your final model, evaluate it on the validation dataset. This can give you a sanity check to see if your estimation of performance has been overly optimistic and has leaked.

Conclusion

Identifying data leakage beforehand and correcting for it is an important part of improving the definition of a machine learning problem. Many forms of leakage are subtle and are best detected by trying to extract features and train state-of-the-art algorithms on the problem. Data leakage and data wrangling are been processed to detect and avoid from further process in healthcare in near future.

References

[1] Fu, X., Gao, Y., Luo, B., Du, X. and Guizani, M. Security Threats to Hadoop: Data Leakage Attacks and Investigation. *IEEE Network*, *31*(2), pp.67-7, 2017.

[2] McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012

[3] Kandel, Sean, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. "Research directions in data wrangling: Visualizations and transformations for usable and credible data." *Information Visualization* 10, no. 4 (2011): 271-288.

[4] Goldston, David. "Big data: Data wrangling." *Nature News* 455, no. 7209 (2008): 15-15

[5] Terrizzano, Ignacio G., Peter M. Schwarz, Mary Roth, and John E. Colino. "Data Wrangling: The Challenging Yourney from the Wild to the Lake." In *CIDR*. 2015.

[6] Endel, Florian, and Harald Piringer. "Data Wrangling: Making data useful again." *IFAC-PapersOnLine* 48, no. 1 (2015): 111-112.

[7] Papadimitriou, Panagiotis, and Hector Garcia-Molina. "Data leakage detection." *IEEE Transactions on knowledge and data engineering* 23, no. 1 (2011): 51-63.

[8] Schouten, Pieter. "Big data in health care: solving provider revenue leakage with advanced analytics." *Healthcare Financial Management* 67, no. 2 (2013): 40-43.

[9] Rauscher, Richard, and Raj Acharya. "A network security architecture to reduce the risk of data leakage for health care organizations." In *e-Health Networking, Applications and Services (Healthcom), 2014 IEEE 16th International Conference on*, pp. 231-236. IEEE, 2014.

[10] Saravanan.N, Pavithra.K, Nandhini.C, "Iris Based E-Voting System Using Aadhar Database",International Journal of Scientific & Engineering Research, Volume 8, Issue 4, pp. 62-64,Apr. 2017.

[11] Jeyakumar, Balajee, MA Saleem Durai, and Daphne Lopez. "Case Studies in Amalgamation of Deep Learning and Big Data." In HCI Challenges and Privacy Preservation in Big Data Security, pp. 159-174. IGI Global, 2018.

[12] Kamalakannan, S. "G., Balajee, J., Srinivasa Raghavan.,"Superior content-based video retrieval system according to query image"." International Journal of Applied Engineering Research 10, no. 3 (2015): 7951-7957.

[13] Ranjith, D., J. Balajee, and C. Kumar. "In premises of cloud computing and models." International Journal of Pharmacy and Technology 8, no. 3 (2016): 4685-4695.

[14] Sikender Mohsienuddin Mohammad,Surya Lakshmisri , "SECURITY AUTOMATION IN INFORMATION TECHNOLOGY", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.6, Issue 2, pp.901-905, June 2018, Available at :http://www.ijcrt.org/papers/IJCRT1133434.pdf

[15] R.R. Nadikattu. 2017. ARTIFICIAL INTELLIGENCE IN CARDIAC MANAGEMENT. International Journal of Creative Research Thoughts, Volume 5, Issue 3, 930-938.

[16] Sikender Mohsienuddin Mohammad, "IMPROVE SOFTWARE QUALITY THROUGH PRACTICING DEVOPS AUTOMATION", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.6, Issue 1, pp.251-256, March 2018, Available at :http://www.ijcrt.org/papers/IJCRT1133482.pdf

[17] Ushapreethi P, Balajee Jeyakumar and BalaKrishnan P, Action Recongnition in Video Survillance Using Hipi and Map Reducing Model,International Journal of Mechanical Engineering and Technology 8(11), 2017,pp. 368–375.

[18] Rahul Reddy Nadikattu, 2014. Content analysis of American & Indian Comics on Instagram using Machine learning", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.2, Issue 3, pp.86-103.

[19] Sethumadahavi R Balajee J "Big Data Deep Learning in Healthcare for Electronic Health Records," International Scientific Research Organization Journal, vol. 2, Issue 2, pp. 31–35, Jul. 2017.