Deep Learning Techniques to Address Issues in Data Quality and Data Variety

C.Pabitha, B.Vanathi

***Abstract:** Deep Learning and Big Data Analytics are key focus in current rapidly growing environment. The use of large data has become crucial to different organizations as they collecting huge amount of domain-specific data, which contains critical information about cyber security, theft detection, national resources, business economics, marketing, and medical information. The assessment of this huge amount of data needs advanced and improved analytical techniques for surveying and guessing future courses of action by making advanced decision-making strategies. Deep learning algorithms utilize the collected training data, to create a representation model. This model uses the computer for predictions or decision making about new data without needing to train the machine explicitly to perform user task. These techniques and algorithms infer greater level complicated abstractions as data are represented through tree like structure. A major use of Deep Learning is processing, learning and training from the huge amounts of unsupervised data, analyze patterns from the data and can be used for large Datasets in which the raw data is largely unlabeled and not classified. In this paper, Deep Learning techniques for addressing Data of different variety/formats is analyzed, enabling fast and full processing and integration of large amounts of different variety of information i.e. Data transformation is also addressed. It also addresses the quality of data as the performances of a machine improve depending on the data quality. Further exploration on the deep learning techniques to assist Big Data by focusing on two key topics: (1) is it possible for Deep Learning to assist some of the specific problems like Data Variety and Data Quality in Big Data Analytics, and (2) Whether these techniques can aid in processing the Big Data*

***Keywords**: Deep Learning, Big data Analytics, Data Transformation, Data Quality, Statistical Models..*

I. INTRODUCTION

Traditional Big data quality analysis methodologies are mainly based on feedback and customer behavior or previously developed business and development strategies. These analysis mainly restricts the working and performance as it is highly time consuming process with very low desirable accuracy. By implementing deep learning techniques, it is possible to leverage resources and methodologies to overcome these limitations and provide greater benefit to users. In this paper, first review of relevant works and analysis of machine learning techniques, tools utilized, and statistical methods are clearly explained. Next, a data quality framework is used depending on the deep learning and statistical methods for identifying quality of data are explained. Developing a deep understanding of the different varieties of data, is an important prerequisite for Data Analysis and Feature Engineering for Deep Learning models... Data transformation means taking data stored in one format and transforming it to another. There are several basic types of data from a Deep Learning perspective they are categorical data, numerical data, time series data, and text data. This can be converted to single format for fast information retrieval and speed data processing. Data transformation a more important and sophisticated role in big data analytics because when dealing with large volumes of data, different types of data analytics tools and different data storage systems situations arises where a large amount of data needs to be transformed from one format to another. One of them major tasks for data quality is to identify problems related to it, specifically outlier detection. A theoretical data quality framework developed by Dr.Richard Wang, MIT professor, represents intrinsic, contextual, and easily accessible framework to address quality [1]. Major data profiling tools contain many data profiling modules, and are based on statistics algorithms, In addition, these tools do not offer insight to increase the data quality problems identified by deep learning algorithms. Deep learning is a popular subfield of artificial intelligence (AI), and is used in self-driving car, data mining, and natural speech recognition, Computer vision etc. [2-5].

II. FOUNDATIONS IN IMPROVING DATA QUALITYAND VARIETY

Most of the research in the field of Data Quality is from communities of Database management. Data quality can be defines according to different data domain aware. In general, it can be observed that data quality is dependent on the quality and originality of the data source [6]. Several definitions can be explained 1) Definition: it is observed that Data Quality is based on the context, domain, or the areas in which it is used. The different variety in big data adopts advanced data structure which raises the complexity of its quality evaluation.

Revised Manuscript Received on November 15, 2019

Ms.C.Pabitha, Assistant Professor, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Chennai, Tamil Nadu, India. Email: pabithac.cse@valliammai.co.in

Dr.B.Vanathi, Professor and Head, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Chennai, Tamil Nadu, India.. Email: vanathib.cse@valliammai.co.in

A. Outlier Detection Technologies

Outlier data is defined as the variety data that is entirely different from the other data available. Hawkins globally defined it as “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”[7]. Anomaly or outlier data does not mean error data but it may lead to some risk while accessing the quality features. Outlier detections are mainly used as applications in several organization and industries. Outlier Detection system, are mainly based on statistical and machine learning techniques, which automatically builds behavioral models that describes the constituents of "normal" behavior of sample entities. These models are mainly based on historical data and are assessed with ongoing observed activities [8]. Statistical models are mainly utilized for outlier detection, which includes Zipf distribution and other normal distributions. In addition these models work well in one-dimensional and multidimensional data

B. Statistical process control (SPC)

SPC [9-11] is a procedure of quality management, and it is based on statistical techniques. In case of normal distribution, it is considered to be sequence of continuous, symmetric, bell-shaped distribution of a variable. Fig.1 represents a probability distribution techniques for quality features with 3 to 6 standard deviations (σ) and mean (μ)

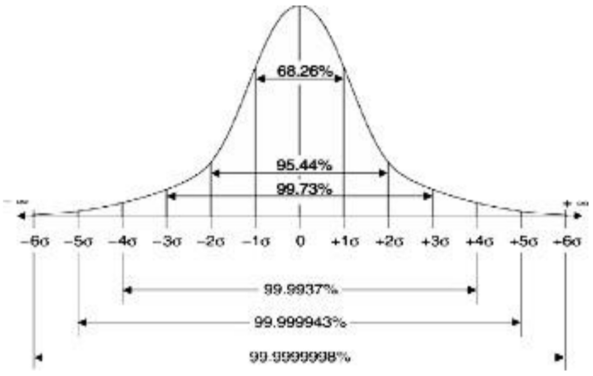


Fig. 1 Normal Distribution Curve

III. OVERVIEW OF BIG DATA

By stating several references such as: [14], [15], [16], [17][18]. Big Data is defined as Big-volume, very high-velocity and variety information that require mostly cost-effective, data processing for improved decision making. The word “Big Data” is used to define a massive volume of both unstructured and structured data; so it's very difficult to implement by using traditional repositories, algorithms and software technologies. **Data Transformation** plays key primary role in big data analytics as it is deals with large volumes of data, several varieties of data analytics tools is likely to encounter situations where a massive amount of data needs to be transformed from one format to another. The Life cycle of Big Data and characteristics are discussed in the following topics.

A. Lifecycle of Big Data management:

- *Data Inception :* In this phase data is generated from several resources that are responsible for the creation of data: audio/video signals, sensors and other sources are used to gather critical information, surveillance, live streaming data from social networks and still images, transaction and other database records, are some of the few to mention.
- *Data Acquisition:* contains collection of data and also data transmission, data pre-processing [12][13].
- *Data Collection:* Data is collected in fixed data formats from several sources such as: real time data measurements using RFID and sensors, or data from any other resources that uses predefined script to crawl and retrieve the web.
- *Data Transport:* To transport the gathered data into different data centers using multiple interconnected networks.
- *Data Pre-Processing:* It contains several preprocessing activities such as Data Integration, Data Enrichment, Data Transformation, Data Reduction, and Data Cleansing.
- *Data Storage:* In this phase, it details about the infrastructure or the center for storing data and distributed among different clusters, present across the world. This systems ensure many fault tolerance levels to attain reliable and efficient storage.
- *Data Processing & Analytics:* In this phase application of Deep Learning algorithm and Data Mining, Artificial Intelligence is used to process the data and gather useful information for better decision making.
- *Data Visualization:* Visualizing the data is the best way of examining the value of processed data and taking decision accordingly. Several algorithms impart the data to be visualized according to the user requirement Visualization method application in Big Data is of an importance as it closes the loop value chain.

B. Features of Big Data

[14], [15] explains in detail the various characteristics and features of Big Data. Volume, Variety, and Velocity are the three main characteristics, to which value, veracity and variability are added. These features provides several facilities which can be used in many recent applications. The amount of data generated by all industries are very huge and are still flourishing rapidly. The data should be gathered, preprocessed before any implementation process. The following Fig2 represents the characteristics of Big Data, which is used in many current real world applications. Also Business enterprises need the data to be rightly analyzed to manage the tough competitions.

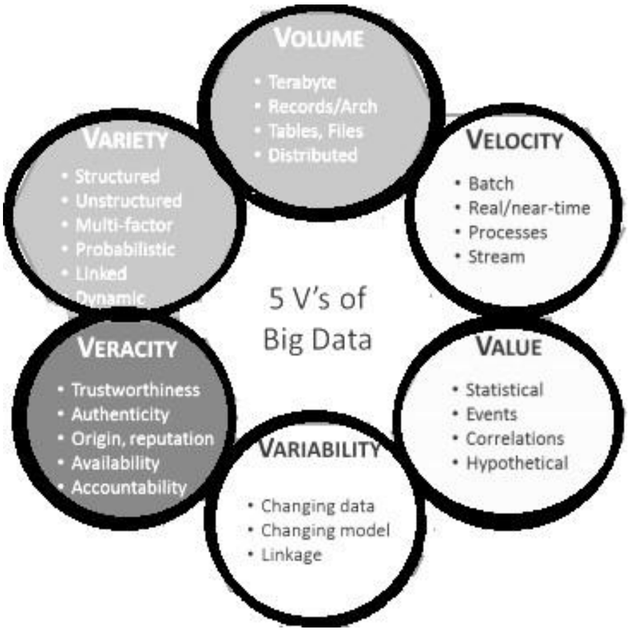


Fig 2 Big Data Characteristics

IV. DATA QUALITY EVALUATION

Evaluation of Big Data Quality mainly depends on lifecycle management of Big Data as it mainly redefines the way data management is processed. Managing the data and improving the quality involves adding more computations in each stage. With quality control being the main focus and monitoring is done to avoid quality failure during several phases of the lifecycle. Quality evaluation of Big Data is majorly about properties that focus on performance, value and cost. The quality of data is measured by six main characteristics that defines the Big data.

A. Six dimensions of data quality

- **Completeness**: Data completeness is the fullness of expectations. Data is considered to be complete if it meets the expected expectations.
- **Consistency:** Data is consistent if all the machine across the organization/world reflect the same information.
- **Accuracy**: Data is accurate if the degree in which data correctly reflects the 'real world' object.
- **Timelessness**: It refers data should be available when required
- **Validity:** Data is said to be valid if it conforms to a particular type, format and range of definition
- **Uniqueness**: data entered should be unique in nature.

Big Data Quality Issues is mainly due to the quality requirements that are not met on data values [21]. These major issues are due to several factors at different life cycle of Big Data:

1) The several data sources fetched are: data copying, unreliability, trust, different sources, inconsistency and data domain, 2) Generation level: Data are generated both manually, readings in sensors devices, social networks, unstructured data, and missing values, and 3) Application level (acquisition: collection, transmission). The data pre-processing techniques improves data quality by executing several tasks and activities like data transformation, integration, fusion, and normalization.

V. DATA TRANSFORMATION

A. Data Transformation and Character encoding

Character encoding problems are maim reason for data transformation. It mainly lets computers to represent singular letters as machine understandable codes. In other words, every letter in the alphabet are represented as a machine code. As long as the machine can able to understand which specific code is used to denote the letters in any given set of documents, it can able to decode the data so that it can translate into numbers and letters and display them on screen. The main problem that arises in character encoding is, the application is used to encode the characters using any format of encoding scheme, and it also sends the same information to any other application that uses an entirely different encoding scheme by default. When the other application tries to manipulate the data, it may not be able to read the characters. In current computing world, most computers follow the UTF-8 encoding scheme, in order to overcome encoding issues.

B. Data Transformation from CSV to XML

CSV also termed as Comma-separated values and XML Extensible markup language are the default ways of storing any data. In CSV file, usage of commas to separate variety of data values from one another. In Extensible markup languages, it represents the data using tags that can define several varieties of data and their corresponding values. As XML and CSV work in very different formats, an application which was programmed to read and write data in XML format won't be able to open a CSV file, and vice-versa. To avoid this confusion data transformation is used.

C. Speech – to – Text and Data Transformation

Human speech can be stored in an audio file, and can be translated into a text file. The pictorial representation of the audio wave is represented in the following Fig 3. Speech/ audio are unstructured data formats that has several preprocessing procedures which is to be before it can be transformed. The first procedure is to upload the data into a machine code or binary format. For this, first take values after every specific time interval.

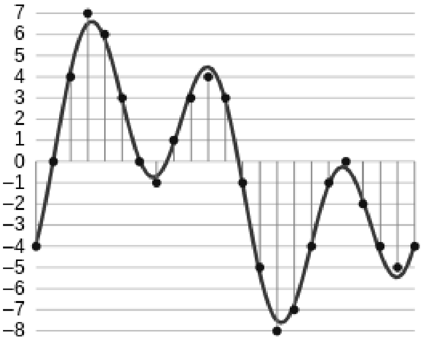


Fig 3 Pictorial Representation of audio wave

For example; when considering two second audio file, we have to extract values at every half a second. This is called **sampling,** and the rate at which it is sampled is called the **sampling rate.** The second method to represent audio data is to convert into frequency domain. Upon sampling an audio data, more data points are required to represent the entire data and moreover, the rate of sampling should be greater as possible. Also, when representing audio data in **frequency domain**, very low computational space is used.

VI. DEEP LEARNING ALGORITHMS

Deep Learning enables the role of data organizer by flagging the data score based on the ratings of probabilistic value that is retrieved from training set of past data organizer and the decisions made and also by categorizing duplicate, missing, incorrect entries. This reduces human effort and managing activities. This process is always useful or the organization to enable a good hold of what type of data is produced and consumed. Deep Learning algorithms aid in deriving quality index of data and the score to assess the quality of data set and reliability in present real time parameter values. These algorithms can also predict trends and gather outliers if trained properly and can also provide valuable suggestions or take actions. Deep learning algorithms can also learn from labels made from human decision on the training datasets and duplicate the scenarios in real-time.

This research paper mainly focuses to provide an overview of the application of deep learning methods to improve data quality. Two main types of supervised learning algorithms are used to increase the quality of data. The first algorithm creates a functional form based on the assumption and relationship between the features and the outcome. Major examples of these algorithms are linear regressions and Logistic regressions. The second algorithm is flexible and does not require any structural assumptions. Examples are random forests, decision trees, and the K- nearest-neighbors algorithm. Decision tree and random forest algorithm are used for the implementation and it was found that random forest provides more accurate result when compared with decision tree.

To illustrate how these supervised learning algorithms work in improving data quality, it proceeds in two steps. In the initial step, the algorithms uses the training dataset to understand the relationship between features and human decisions. In the later step, the algorithms applies the decision taken in the first step to validate the dataset and suggest predictions of the probability that how data quality can be improved. When comparing the performance of the Logistic and the random forest algorithm. It is analyzed that random forest algorithm provides a large improvement in accuracy.

A. Linear Regression

Regression is a way of modeling the end value based on several independent values. This methodology is used for forecasting the value and finding out the causes its effect and the relationship between these variables. Regression techniques usually differ based on the count of independent variables and type of relationship found between independent and dependent variables.

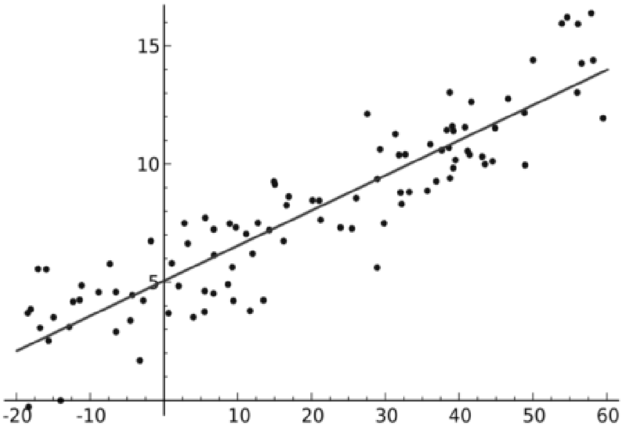


Fig 4 Linear Regression

Linear regression as represented in the above Fig 4, is performed by predicting any dependent variable value (y) which is based on an independent variable (x). So, a linear relationship works out between x (input) and y (output).

B. Logistic Regression

Logistic Regression is a type of statistical analysis (also termed as *logit model*) is mainly used for predictive analytics and modeling, it can be extended to applications in deep learning. Logistic Regression is the most general and popular algorithm used to solve any classification problem. In this approach, the dependent variable is discrete or categorical: either A or B also termed as binary regression or a range of finite sequence options like A, B, C or D as multinomial regression. It is utilized in statistical software so that understanding the relationship between one or more independent variable and dependent variables is done by estimating probabilities using a logistic regression equation. This analysis help in predicting the likelihood of outlier data among the data set. It classifies the data as good data and bad data by using a sigmoid function which helps in categorizing the data and enables the quality of data to be identified. Logistic regression is mainly used to produce estimates of the intercept made from the model and coefficients, this can be applied together with quality statistics for every individual parameter and also for the model as a whole. When Logistic regression is applied to new data, it provides a probability that ranges from zero to one which reflects the relative likelihood that the data belongs to the good or bad category, upon giving the known values of predictions. Fig 5 represents an example of Logistic regression technique.

The most generally used algorithm to estimate logistic regression models is the maximum likelihood algorithm. Maximum likelihood is a recursive, iterative algorithm; which assigns the initial values to the model coefficients, and tests the solution with the training data, the model is improved by repeated iteration and testing until no further improvement or iterations are needed.

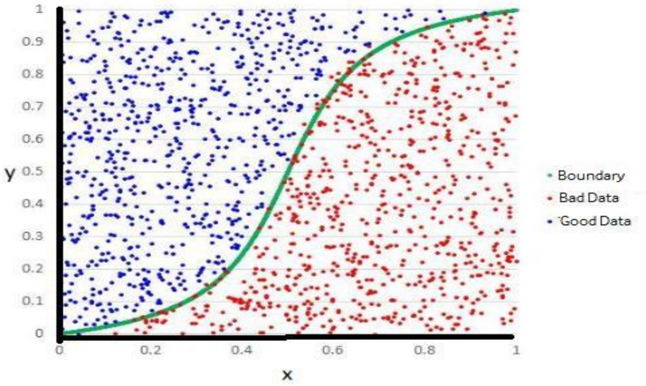


Fig 5 Representation of Logistic Regression

C. Random Forest Algorithm

Random forest involves building several decision trees by combining the output thus improving the ability of the model to generalize. It comes under the category of **Ensemble learning** is a supervised learning method that combines the prediction of several smaller models thereby improving predictions and generalization. Ensemble methods result in a significant higher performance than any of the individual base models like decision tree can achieve.

1) Bias-Variance Tradeoff

Bias and Variance are the two source of error that restricts models from generalizing beyond the given training set. Bias is for simplifying the assumptions made by the model to understand the learning easier. Models like linear regression and logit regression have underlying assumptions and are considered to have the high bias, whereas some other models like K nearest neighbors and Decision Trees are considered to have low bias. Variance is to find how much the model will differ if several different training data were used. If a model is considered to be having higher variance, then it means that the model has learned with training data supplied to it. Models such as linear and logistic regression have low variance and models such as Decision Trees is considered to have high variance. Any model used, the main objective is to reduce both bias and variance errors. Therefore it is necessary for these two source of errors to be balanced so that the model produces more reliable results.

2) Decision tree vs Random Forest

A Decision tree (refer Fig 6) performs in the following steps:

- Provided a data set (n x p), a tree separates the data based on if-else condition. A tree creates inferences as rules and the rules partition the data set into several distinct, non-overlapping regions. These rules are identified by a variable's participation to the homogeneity of the resulting child nodes (X2, X3).

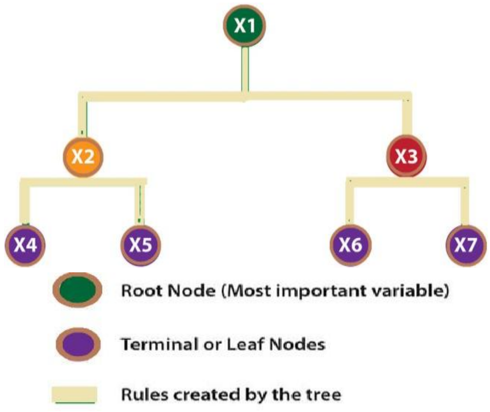

X1
X2
X3
X4
X5
X6
X7
Root Node (Most important variable)
Terminal or Leaf Nodes
Rules created by the tree

Fig 6 Decision tree representation

- Since Variable X1 is the top most homogeneity in child nodes, it automatically becomes the root node. An identifier at root node is said to be the important variable in that particular data set.

To decide at what point the tree should be split the following points are used.

- The output of a Regression tree is predicted by means of observing the target node. The way of splitting the tree is based on minimizing RSS (Residual Sum of Squares). The node or the variable which proceeds to the largest reduction in RSS is suggested and chosen as the root node. A **top- down greedy** approach, methodology is taken for splitting the tree and this is also termed as ***recursive binary splitting***.

Under **classification of trees** in which the target is predicted by observing the terminal node, the splitting of decision which is done on the following steps:

- **Gini Index** – Used to measure the purity of node. If the index value takes the least one, it informs that the node is pure. For splitting, the Gini index of a child node should be way less than the parent node.
- **Entropy** - For measuring node impurity Entropy is used. Consider a binary class (a,b), Now the formula to calculate entropy is given as follows. Entropy is considered to be maximum at p = 0.5. For any p(X=a)=0.5 or p(X=b)=0.5 it means that, a new observation has a chance of 50%-50% getting classified in any one of the classes. The entropy is considered to be minimum if the probability is 0 or 1. Fig 7 represents the entropy diagram.

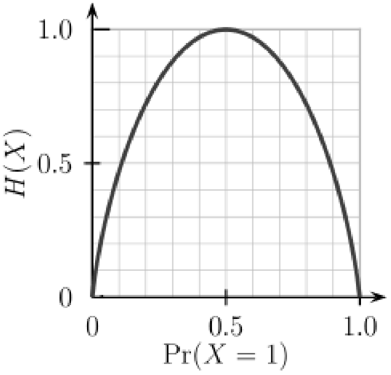


Fig 7 Entropy

Entropy = - p(a)*log(p(a)) - p(b)*log(p(b))

Every tree attempts to generate rules so that the resultant terminal nodes should be very pure. Higher the value of purity, lesser the decision due to uncertainty. But generally decision tree suffers high variance, means getting greater prediction error on missing data. This can be overcome by using more data for training. But the data set available for processing is very much limited so it is necessary to use the re sampling techniques such as bagging and **random forest t**o generate more data.

Combining several **decision trees** together will results in a **Random forest**. Working of Random Forest is as follows:

- In first step Bagging is used, it is a Bootstrap Aggregating algorithm that is used to generate random samples. Provided a simple data set S1 (x rows and y columns), create a new dataset S2 by sampling x cases at random. Replacements are made from original data. About 1/3 of the n rows from S1 are left out known as Out of Bag(OB)samples.
- Second step the method trains on S2. OB sample is now utilized to generate unbiased estimation of the error.
- Out of the y columns, Y << y columns are gathered at every node in the data set. The Y columns are collected at random basis. Formally, the fixed choice of Y is y/3 for the regression tree and Y is sqrt(y) for the classification tree.
- Unlike tree, No pruning takes place differently in the random forest; i.e., every tree is grown fully. In the case of decision trees, pruning is always used to avoid over fitting. It can be defined as a set of sub tree that proceeds to the least test error rate. Instead of that usage of cross validation is done to find the possible test error rate of a sub tree.
- The process is repeated to grow several trees and the final decision is obtained by voting or averaging.

VII. IMPLEMENTATION

A. Tensor Flow Data Representation

One of the most popular deep learning library namely **Tensorflow** is used widely to experiment these learning algorithms. As tensor flow is mainly based on data flow graphs that is used for building models. The Tensorflow library helps in creating neural networks with several layers. Tensors are nothing but a representation of the data in deep learning.

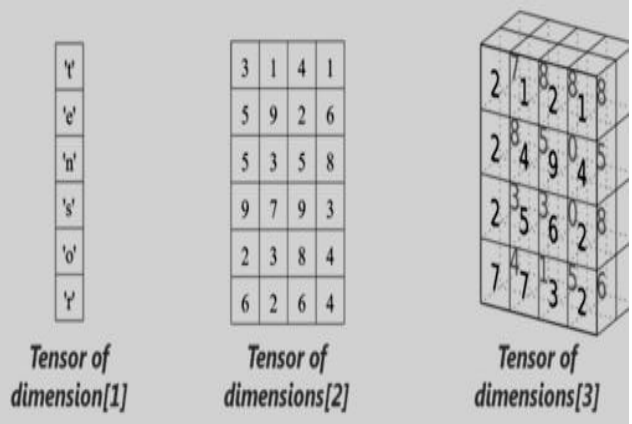


Fig 8 Multi-dimensional array represented as tensors

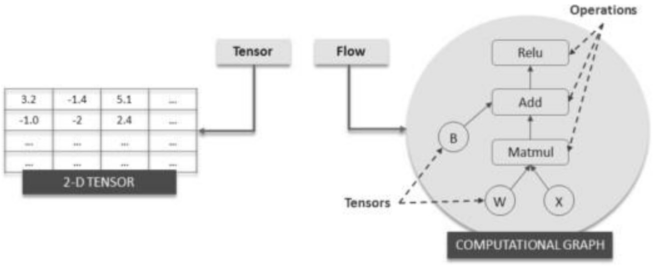


Fig 9 working of tensor flow

As displayed in the above image, tensors can be represented as multidimensional arrays, which can be used to denote data having greater dimensions. Generally, Deep Learning uses high dimensional data sets in which the dimensions represent several parameters in the data set.

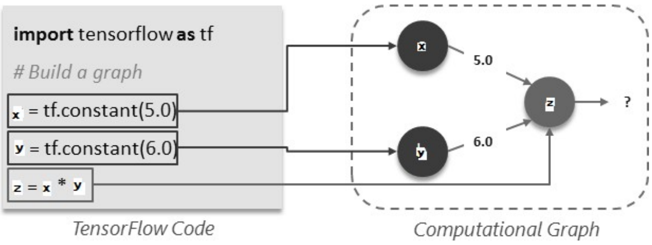
TensorFlow is a python library that generates different varieties of functionalities for experimenting Deep Learning Models. Based on our earlier discussion, the term TensorFlow is a combination of two terms – Tensor & Flow: the name tensor is used to denote data in multi- dimensional array format and the word flow denotes the several series of operations that is performed on tensors as shown in the above image. The figure represents the two step processing using tensorflow.

Tensor Flow program is a two steps process:

1. Constructing Computational Graph
2. **2.** Implementing Computational Graph

1) Constructing Computational Graph

Computational graphs are considered to be sequence of TensorFlow methods arranged as different nodes in graph. Every node takes 0 or many tensors as input and provides a tensor node as output. For example consider x,y & z nodes represented as computation graph as follows in Fig 10.



```
import tensorflow as tf
# Build a graph
x = tf.constant(5.0)
y = tf.constant(6.0)
z = x * y
```

Fig 10 computational graph in tensor flow

2) *Implementing computational graph*

To implement the above example

```
import tensorflow as tf
# Generate a graph
x= tf.constant(5.0)
 y = tf.constant(6.0)
z = x * y
```

B. *Tensor Flow Data Validation*

TensorFlow Data Validation is used to identify the anomalies or outliers in training and serving data, it can also create a schema automatically by examining the data. This can be configured to identify different classes of outliers in the data. Tensor flow is also used to perform quality check by comparing data against a schema that confides to the expectations of the user.

It compares the example in training and serving data and detects the bad data. For implementing the logistic regression using tensorflow import the necessary libraries. Using Numpy along with Tensorflow for basic computations, Pandas for Data Analysis and Matplotlib for plotting the data helps in classifying the data. Further using Scikit Learn for preprocessing module and Encoding the data is performed. Using these libraries and implementing the decision tree and random forest algorithm for categorizing good data and bad data proves that random forest algorithm provides better result when compared with decision tree.

VIII. CONCLUSION AND FUTURE DIRECTIONS

Improving Data Quality is considered as one of the most challenging issues in the era of Big Data. Present techniques and solutions are mainly from academia and industry that uses data in daily basis and quality matters a lot. Evaluating and assessing the quality of Big Data is of greater importance to government organizations and industries. In addition, proceeding with well explored data quality plan, using the right assessment tools and adopting the already available quality measurement approaches, thereby conducting several quality evaluation activities all result together in achieving high quality assessment results. Data transformation to a common machine understandable code not only simplifies the job but also improves the productivity to a greater extent. Further, addressing quality and data transformation across the Big Data enforces a wider use among different customers and retailers. Finally, it also guarantees a cost-effective quality management processes. The paper addresses some deep learning algorithms that greatly improves the data quality providing better result than the existing algorithms. A different feature weighing method and tree selection method can be used to further improve the classification performance and random sampling.

REFERENCES

1. Diane M Strong , Yang W.Lee, Richard Y.Wang " Data Quality in Context" Communications of the ACM Volume 40, Issue 5, May 1997, Page 103-110
2. Martin Hofmann1 , Florian Neukart2, 3, Thomas Bäck, Artificial Intelligence and Data Science in the Automotive Industry January 2016
3. I. Lighthill, "Artificial Intelligence: A General Survey," in Artificial Intelligence: A Paper Symposium. London: Science Research Council, 1973.
4. S. J. Russell and P. Norvig, Artificial intelligence: a modern approach (3rd edition): Prentice Hall, 2009.
5. M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," The knowledge engineering review, vol. 10, pp. 115-152, 1995.
6. Gary Mazzaferro, BIG DATA General Reference Architecture, November 2011.
7. Tung Kieu ; Bin Yang ; Christian S. Jensen, "Outlier detection for multidimensional time series using Deep Neural Networks" IEEE International Conference on Mobile Data Management. June 2018
8. A. C. Atkinson. Fast very robust methods for the detection of multiple outliers. Journal of the American Statistical Association, 89:1329–1339, 1994.
9. Maria Emilia Camargo ; Walter Priesnitz Filho ; Angela Isabel dos Santos Dullius "Statistical quality control: A case study research 4th IEEE International Conference on Management of Innovation and Technology 2008
10. kai yang & walton M. hancock "Statistical Quality Control For Correlated Samples" May 1989
11. William H. Woodall Statistical Quality Design and Control, Second Edition, Journal of Quality Technology, 39(1), pp. 88–89, Volume 35, 1993 - Issue 3, Mar 2012
12. Han Hu ; Yonggang Wen ; Tat-Seng Chua ; Xuelong Li "Toward Scalable Systems for Big l," *IEEE Access*, vol. 2, pp. 652–687,2014.
13. M. D. Anto Praveena ; B. Bharathi A survey paper on big data analytics International Conference on Information Communication and Embedded Systems (ICICES) Feb. 2017
14. D. P. Acharjya, Kauser Ahmed P, A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, (IJACSA)International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.
15. A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015), pp.137-144.
16. D. R. Luna, J.C Mayan, M.J. García, A.A. Almerares, and M. Househ "Challenges and Potential Solutions for Big Data Implementations in Developing Countries

AUTHORS PROFILE

C.Pabitha.,M.E.,(Ph.D), Assistant Professor of Department of Computer Science and Engineering, SRM Valliammai Engineering College has more than ten years of teaching experience in Academic institutions. She has published more than twenty research papers in reputed International Journals and Conferences. She has authored two Books on Python Programming and Multi Core Architecture. She has organized many workshops, seminars, Entrepreneurship awareness camp funded by DST-NIMAT and FDPs. She is an active member in CSI, ACM, ISTE, IEEE and Indian Science Congress. Her area of interest include Big Data, Artificial Intelligence, Deep Learning, and Computer Vision.

Dr.B.Vanathi, Professor and Head, Department of Computer Science and Engineering, SRM Valliammai Engineering College has more than 22 years of teaching experience. She completed her Ph.D. at Anna University in the year 2013. She has published more than 50 International Journals and Conferences. Her research area includes Context Aware Computing, Data Mining and Machine Learning. She is an active member in ISTE, CSI, IAENG, IDES, IARC, ACEEE and Indian Science Congress. She has received many research grants under DST, SERB and AICTE. She is currently guiding five Ph.D research scholars.