Data and Pre-Processing of Data For Machine Learning

Sriram Gupta Kaluva[1], R. Jaya[2]

Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore-560103 Karnataka, India

[1]kaluvasriram@gmail.com
[2]jayamanojkumar@gmail.com

***Abstract*— Machine learning is one of the powerful tools for gleaning knowledge from massive amounts of data. While everyone is focused on improving the efficiency and accuracy of the algorithms available, there is less attention given to the equally important aspect of monitoring the data fed to the training algorithms. The importance of this topic is hard to dispute: One of the main problems is that any of the errors in the data fed to the algorithms will lead to false predictions and inaccuracy of the algorithm. This leads us to a situation which treats training the algorithm and the served data as an important production asset**

***Keywords*— Machine learning, Data, Algorithm, Accuracy, Big data, Pre-processing.**

I.INTRODUCTION

This is an era of "Big data", where the world is dependent on the data produced by various sources available including from the transactions made by the user to the tweets made in twitter. In the 1900s only big companies had access to the data and it was only useful to them, but now the data usage has become more prominent, that a person with minimal knowledge on machine learning is able to use it properly. Imagine a scenario in which -if the data available is such prominent and what if the trueness of the data that the developer had received is overseen and blindly followed in research fields. This can nullify any benefits on speed and accuracy for training and inference.

A. BigData

Big Data has 4 important characteristics, namely Volume, Variety, Velocity, Complexity. As organisers collect a large amount of data from various sources, this proves the Volume characteristic of big data. Data streams at an unprecedented speed, it must be dealt with in a timely manner. This proves the Velocity of Big data. Data we gather comes in all types of formats such as structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. This shows the Variety of Big data. Today's data is coming in different formats and from different sources. So it is hard to link, match, cleanse and transform the data across systems. This shows the Complexity of Big data.

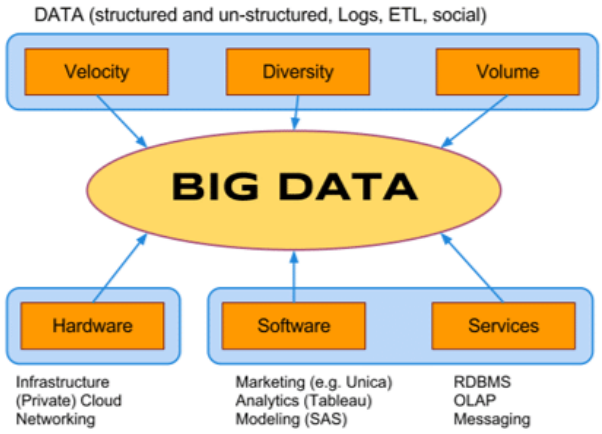

Fig.1 Big data

The above figure shows the general representation of big data.

B. Data Flow in Machine learning

Machine learning is purely based on the data (Also called as a training set(includes test set)) that is needed for the project. Most of the machine learning algorithms Pre-process the data using some pre-processing techniques and gets useful parameters from the data. Then using algorithms like 'Feature selection algorithm', the useful features are extracted (called as feature splitting). So from the extracted step, the user can add more features to the data which he/she feels important. After the data is extracted properly, it is divided into two different data sets which are training set and test set. The training set is fed to the algorithm to create a training model. And the test set is used to test the efficiency of the model that has been created.

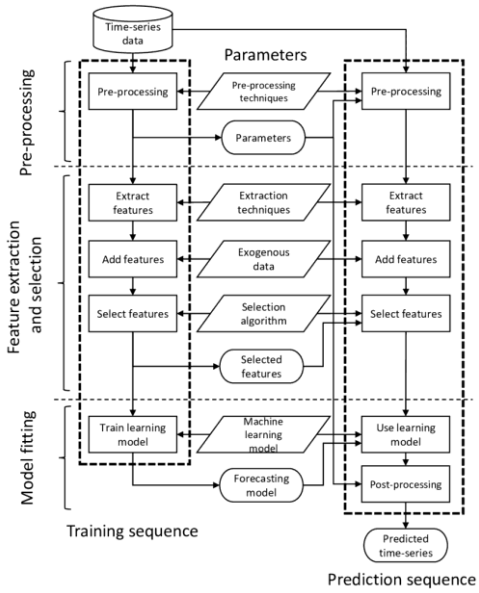

Fig. 2 Data Flow example

The above figure shows the typical example of the data flow in Machine learning.

C. *BigData in Machine learning*

There are Learning algorithms mainly concerned with issues related to Volume and Variety. Machine Learning algorithms deal with massive amounts of data i.e., Volume whereas shallow learning algorithms fail to understand complex data patterns which are inevitably present in large data sets. Moreover, Machine Learning deals with analysing raw data presented in different formats from different sources i.e., Variety in Big Data. This minimises the need for input from human experts to retrieve features from all new data types found in Big Data.

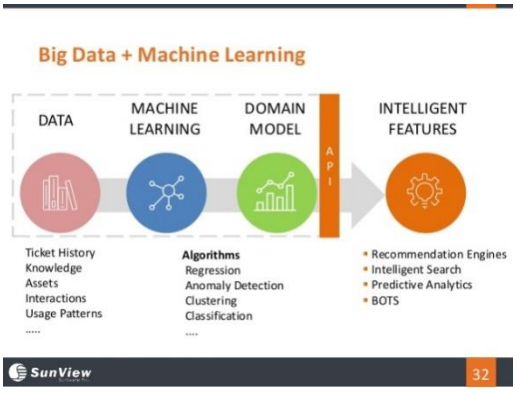

Fig. 3 Big data + Machine Learning

As the businesses are eagerly pursuing big data analytics, it only stands to reason that they'd look for the methods and strategies that will best help them get the most out of it. There are many ways to perform analytics, it completely depends on the business and what insights organisations want to gain or the project that the user is working on. With this kind of variation, Big data growth has increased rapidly. By now almost 75% of companies in the world are making initiatives on Big data and many companies have found it difficult and at times arduous process using big data. Since it is time taking and not accurate to analyse the business data manually, companies came up with a solution to use Machine learning. Machine learning is not new, the concepts of machine learning were used decades before too. But when the big data took off the market analytics, machine learning came into existence again. The main focus of machine learning tends towards the development of algorithms in order to process large amounts of data in real time. The aim is to make predictions or to generalise new input data based on the trends of past data which we got from Big data. As referenced above, big data involves enormous sets of data gathered from a variety of sources. Analysing big data with traditional techniques can only go so far, but with machine learning, predictive information can now be delivered with more accuracy much more quickly. Since there's no need for human intervention in this process, more complex sets of data are now open for big data analysis. So in this way, machine learning offers more accuracy, scale, and speed needed to fully analyse the data that organisations can now collect. And speaking of the future, big data machine learning will likely play an even more integral role in new technologies. Big data has accomplished much already, but machine learning's role will be one of unlocking its full abilities.

II. PRE PROCESSING DATA

A. *Data Standardization*

The data available in the world vary greatly from one organisation to the next since the data is collected from different sources and in different formats and different information models. So, Data Standardization is a process of converting the data into a common format which allows for collaborative research, large-scale analytics, and sharing of sophisticated tools and methodologies. There are so many resources to convert different type of data into a common data model such as the OMOP data model provided by OHDSI. The OMOP data model has a plethora of tools which are used to take advantage of your data model once it is in Common data model format. It allows for the systematic analysis of disparate observational databases. The main goal behind this approach is that it transforms the data contained in the databases into a common format (data models) as well as common representation (terminologies, vocabulary, coding schemes) and then it performs systematic analyses using a library of standard analytic routines that have been written based on the common format.

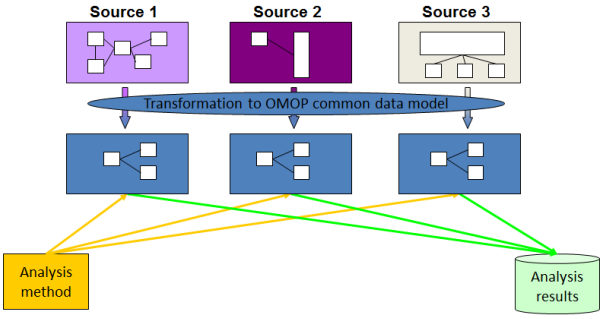

Fig. 4 Flow of OMOP data model

The above figure shows the working of the OMOP data model in which the data is converted from different formats to a common format and analysed.

B. Data Normalization

It is a process in which data attributes within a data model are organized to increase the cohesion of entity types. Simply, it is a process of reducing or eliminating data redundancy in a database as it is incredibly difficult to stores objects in a relational database that maintains the same information in several places. By Data normalisation, the information will be stored in one place and one place only, reducing the possibility of inconsistent data. And by data normalisation, it is easier for object-to-data mapping.

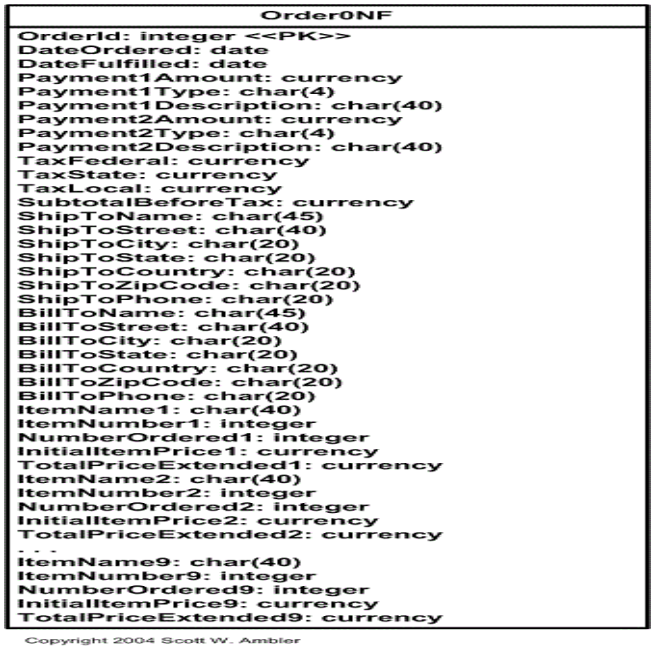

Fig. 5 Initial data base

Normalization has three main steps involved in it. They are:

1. First Normal form: An entity type is said to be in 1NF if it contains no repeating groups of data.
2. Second normal form: For a database to be in 2NF it has to be in 1NF and all of its non-key attributes are fully dependent on its primary key.
3. Third normal form: For a database to be in 3NF it has to in 2NF and all of its attributes are directly dependent on the primary key.

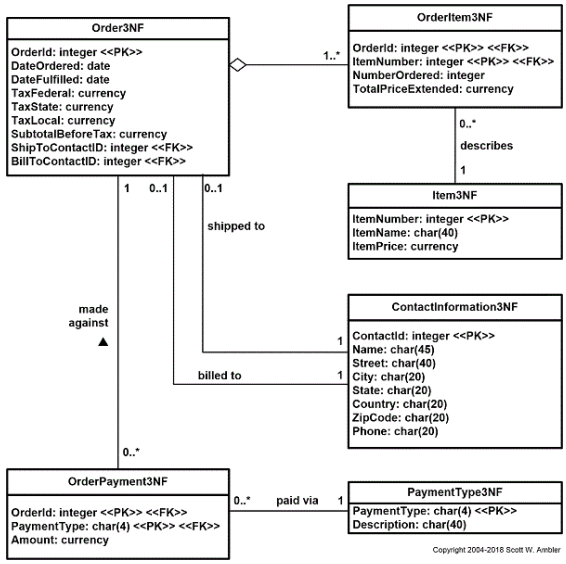

Fig. 6 Normalized data base

C. Data Discretization

Discretization is viewed as the partitioning of a continuous-valued attribute into an ordered discrete attribute with a number of discrete intervals, which equivalent to the process of reducing the number of states of an ordered discrete random variable by combining some of its states together.

Basically, Discretization is a process to transform the range of continuous attributes into a discrete partition which consists of the number of the intervals associated with the boundary set and the quanta set. In inductive learning, a smaller number of intervals are preferred as a larger number means a larger number of possible attribute values which leads to slow and inefficient learning.

The general discretization methods are

1. Class-Attribute Dependent Discretizer (CADD).
2. Maximum Entropy (ME).
3. Equal Information Gain (EIG).
4. Equal Interval Width (EIW).

D. Principal Component Analysis

It is a statistical procedure which uses an orthogonal transformation that converts a set of correlated variables to a set of linearly uncorrelated variables called principal components. It is a tool which is being used most widely by many exploratory data analysis and in machine learning for predictive models. Basically, it is an unsupervised statistical technique that is used to examine the interrelations among a set of variables. It is also called as a general factor analysis which deals with regression in determining a line of best fit. Let's say we have n number of observations having p variables, then the number of distinct principal components is min(n-1,p). It is defined in such a way that the first principal component has the largest variance and every succeeding component, in turn, has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vector is an uncorrelated orthogonal basis set.

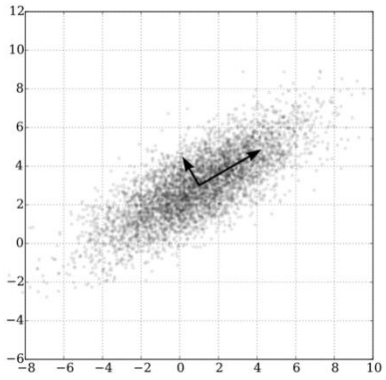

Fig.7PCA of a Multivariate Gaussian distribution

The above figure is centred at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue and shifted so their tails are at the mean.

III. CONCLUSION

So, It doesn't matter how well the algorithm is designed until unless the data is large and preprocessed before it is fed to the algorithm. Preprocessing of data depends on what type of data that is available and types of arguments and number of parameters that are being used to train a model. So, the data that is being used and preprocessing of the data plays an importantrole in the accuracy and efficiency of the algorithm that is being used to train the model.

REFERENCES

[1]. EthemAlpaydin, Introduction To Machine Learning, Third-Edition.

[2]. "BIG DATA IN MACHINE LEARNING.", qubol.com, [Online]. Available: https://www.qubole.com/blog/big-data-machine-learning/ .

[3]. "DATA-STANDARDIZATION.", ohdsi.org, [Online]. Available: https://www.ohdsi.org/data-standardization/ .

[4]. "DATA-STANDARDIZATION.", minitab.com, [Online].Available: https://support.minitab.com/en-us/minitab/18/help-and-how-to/calculations-data-generation-and-matrices/standardize/standardize-columns-of-data/ .

[5]. "NORMALIZATION", agiledata.org, [Online]. Available: http://agiledata.org/essays/dataNormalization.html .