Big Data Analytics with Machine Learning: Challenges and Opportunities

1. Dr.ECCLESTON
Assistant Professor
Department of CSE
AVN Institute of Engineering & Technology
Hyderabad, Telangana.
E-Mail:graju.mtech@gmail.com

2.Dr.THOMAS FELDMAN
Assistant Professor
Department of CSE
AVN Institute of Engineering & Technology
Hyderabad, Telangana.
E-Mail: bpannalal555@gmail.com

Abstract:
Nowadays Big Data analytics and machine learning are most used technology for research in several analytics and computations. In several organizations like health, education, banking and other sectors, the importance of the big data analytics and machine learning play a vital role for classification and predication Most of the research areas for pattern reorganization. Machine learning models require data as an input to them; sometimes the more extensive the data, the better are the results of the machine learning model. In such a situation, big data is fed as an input to a machine learning model to produce the desired output. Big data can be one of the input sources for the machine learning model. In this paper, we have been discussed several challenges and Opportunities using big data analytics with machine learning scenario.

Keywords: Machine learning; big data; data preprocessing; evaluation;

I. Introduction

Big data might benefit from ML, although data preparation for data mining/big data is per se a topic. Big data might favor development and tuning of incremental/online/stream-oriented ML algorithms; especially the models already developed for drifting data might be of interest. The learning comes from extensive calculations done over existing datasets to create a learning model (in most cases) [1]. A normal system can't handle very big dataset calculation and data size is increasing day by day, thus the obtained model should be adapted accordingly. To obtain this we have to implement distributed computing using big data technologies like Apache Mahout, Spark, R-Hadoop or initial analytics processing in projects like hive/ pig and feed output to machine learning algorithms for model/ learning generation. Big data techniques are used in machine learning. We all know that machine learning works well with big sets of data and this where big data comes into the picture. We use big data to extract some hidden information or meaningful insights from huge chunks of data. In a nutshell, we can say that without big data, machine learning would be of no use.

Difference between data science, data analytics, data mining, and big data:

Data science, allows you to extract knowledge from raw data, raw data which is but information. It is an interdisciplinary field; i.e., it uses techniques from many fields like mathematics, statistics, data engineering, visualization, data warehousing, etc., with the aim to extract useful knowledge from the available data/information.

Big Data [2], on the other hand, is a collection of huge amount of data that requires a special database management system. This helps in analyzing and drawing useful insights from the available humongous data (1 Terabyte and above in 2018 is considered huge). In this, we usually apply data science techniques and algorithms; so as to gain the said useful insights.

As we can see, Data science is not restricted to big data; but the fact that data is increasing in size with passing time, it is evident that big data is becoming an important aspect of Data Science.

Data science is a growing field, and it is sought-out by many multinational companies today. Due to the lesser number of skilled, proficient data scientists in comparison to the demand; there is a space that needs to be filled.

Data Analytics [2] is a process in which data is examined in order to draw insightful conclusions. This helps Businesses in making decisions which will prove to be profitable, the techniques used in Analytics are the same as those used in Business Analytics and Business Intelligence. In Data Analytics you will be searching for specific answers; i.e. there will be a test hypothesis for which you will try and find specifics.

For accurate Data Analytics, one needs various tools. I found Data Visualization using Tableau to be a good place to start, and needless to say, Python or Rare important programming languages to know. These languages will enable you to perform seamless data analytics.

Data Mining [3] as the name suggests "mines" data using components of artificial intelligence, traditional statistics, etc. Data Mining unlike Data Analytics is performed without any said hypothesis. Data Mining does not aim to answer specific questions.

II. Big data and Its Life Cycle

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

Big data [4] is the tremendous growth and collection of structured as well as unstructured data. Expanding on that-

1. *Volume:* Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.

2. *Velocity*: Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

3. *Variety*: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

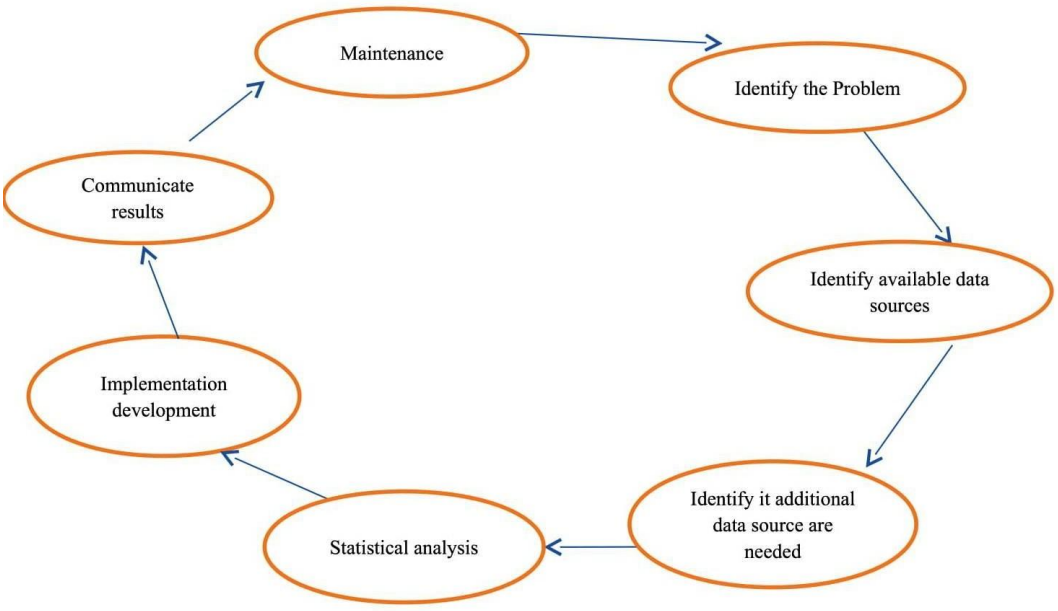

Figure 1.Lifecycle model [4]

So, the 7 steps that together constitute this life-cycle model are:

1. Identify the problem
2. Identify available data sources
3. Identify if additional data sources are needed
4. Statistical analysis
5. Implementation, development
6. Communicate results
7. Maintenance

Data analytics and decision-making are not simple aspects. Some of the major challenges of big data are capture, analysis, data storage, search, sharing, visualization, querying, transfer, update, and information privacy. This involves data sets that are vast, varied, voluminous and complicated, simultaneously. These datasets need to be analyzed using robust and scalable data processing software, under the guidance of a reputed expert like LTI.

Here are a few challenges faced while doing data analytics:

1. The volume of big data is so vast, making it difficult to analyze in a short span of time.
2. Big data has loads of variation, thus requiring the staff to sit and sort it through in order to analyze it.
3. It collects at a high rate, thus making it challenging to keep up with analytics.
4. Data quality is sometimes heterogeneous, resulting in unclean data collection.
5. It is sometimes difficult to generate data insights in a timely manner, with the volume of incoming data.
6. Securing big data is a little difficult, as a lot of licenses and permits are required.

Machine Learning and Its Applications

Machine Learning is an application of Artificial Intelligence and is revolutionizing the way companies do business. ML essentially means a machine which has the capacity to learn. So depending on your past searches, the Amazon or Facebook website code will have learnt what sort of items you are looking for. Thus, when you log in, the site will prop up those items right in front of you. And it may even combine them with an exciting deal. Such strategies are intended to promote impulse online shopping. From what I have heard, Amazon is in the process of implementing or has already implemented a system by which they can even predict the demand for particular items in a specific region and manage the stocks of such items accordingly. Thus, it significantly cuts down delivery time and enhances the Customer Delight.

ML from what I believe is a nascent field, but growing rapidly, and finding applications in increasingly diverse areas. Companies in E-commerce, police departments, sports betting companies, etc. are increasingly hiring Machine Learning specialists to devise intelligent algorithms to enhance the predictive prowess of their systems.

III. Difference between Data analytics and Data mining

The main difference we can observe is that Data Analytics looks for specific answers with a specific hypothesis, whereas, Data Mining does not have a specific answer to fulfil.

Data Analytics has huge scope in **Business Analytics** and Business Intelligence. Data Mining on the other hand uses techniques, both mathematical and scientific to find patterns and trends.

IV. Challenges for machine learning algorithms for big data analysis?

The biggest challenge is to run algorithms across the distributed environment. I don't develop algorithms but I have a hard time even defining the problem even with my technical background. In a nutshell, this is about fitting models in parallel. Tackling this problem means that both data and algorithms scale. Operationally speaking I believe the biggest challenge is that wrangling is a very time consuming process both in development terms and running times. When the data lake has a couple of high volume data sources, it's fine. But when the data lake has

hundreds of data sources the engineering effort to build and maintain the data lake is huge and data engineer's time could be put into much better efforts.

V. Big data is compared to Machine Learning?

Machine Learning and Big Data as such have no direct relation. Although one can say that

Big Data:

i) Big Data Techniques can be used in Machine Learning. Machine Learning usually works with huge chunks of data.

ii) Big Data not only maintains huge amount of data but the real work of big data is analytics where viable and useful information must be extracted from this huge volume of data.

iii) Traditional analytics tools and basic methods are not well suited to capturing the full value of big data.

Machine learning [7]:

i) Machine learning is ideal for exploits and provides insights into the big data.

ii) ML is data driven and runs at machine scale and works on growing datasets,

iii) ML can be used for predictive analytics, knowledge extraction and interpretation and ML focuses on the development of fast and efficient algorithms for real-time processing of data with as a main goal to deliver accurate predictions for example fraud data detection in big data and ML allows mathematical calculation on big data in a much faster way.

iv) ML employs generic methods and real-time and highly scalable predictive analytics so the bottom line is that machine learning is the perfect marriage partner for big data analytic.

Big data analytics and Machine learning are close neighbours who may or may not work together at times to achieve certain goal.

Machine learning is applying mathematics to analyse past data and predict future value or an action. One doesn't necessarily need big data to apply a machine learning algorithm to get a future prediction.

Big data analytics is applying analytical techniques to past data to get some insights and better understanding of data with help of one or another tool in big data environment. This doesn't necessarily need machine learning.

Now, machine learning and big data both can work together in most and all situations to get same or better results.

VI. Big data Analytics: Opportunities with Machine learning

The importance for anything related to data is shooting up. Literally today's data centres occupy an area of land almost equal to 6500 football fields. Literally 100+hours of video uploaded to you tube every minute and it would take 15 years to watch every video. Since data driven decisions yields more result.

Now days, even Iot (Internet of things) generate more data than ever and data is generated rapidly and hence Big data and as said data exploration or predictive modelling is really helpful for getting data insights for any move basis on data, so field of ML is future. In short AI solves the purpose and future can be this…

D**ata Analytics** and **algorithms** are correlated.i.e.To perform analysis we use appropriate algorithms to refine/process the data towards the goal of "what should we get /achieve at the end?"As a solution we need to figure out the *Model* that drives us towards goal. The *Model* here is an Al*gorithmic/Mathematical Model*. Having knowledge in **probability & Statistics**, **linear algebra and calculus** will be a big plus and drives the process of analytics as smoother as possible.

As per my knowledge some critical challenges include:

1) Understanding the Data: We can't do analysis if we don't understand our data in the first place. That means, we should have good knowledge of the type of data we are dealing with, sources of data sets from which they are coming in and what should be derived from the data as the result.

2) Domain Knowledge: Having good knowledge on the domain of which the target data is related to. Understanding the customer's behaviour is a key w.r.t particular domain. Ex: - Retail, Telecommunications, Social Networking, Automobiles, Banking etc.,

3) DataIntegration [5]: The ability to combine data that is not similar in structure or source and to do so quickly and at reasonable cost. With such variety, a related challenge is how to manage and control data quality so that you can meaningfully connect well understood data from your data warehouse with data that is less well understood.

4) Accuracy of Data: This is related to **data quality** issue. As the data comes from different sources, there are at most chances for junk in them as well. We have to ensure that we are processing/ analysing the data of less junkie.

5) Data Volume: Analysing and producing insights to the customer in a timely manner from huge volumes of data is the real challenge.

6) Methods/Tools used: Having Knowledge of right algorithms to be used for processing and analysing the data rather than just using automated tools shows the huge difference in correctness and reliability in a considerable way [6].

VII. Conclusion

In this paper we have been discussed among several research domains such as data science, data analytics, data mining, and big data. Furthermore we have been discussed lifecycle of big data. This paper enlighten the readers and scholars what are the Opportunities Big data Analytics with Machine learning, and how Big data is compared to Machine Learning and what are the Difference between Data analytics and Data mining, finally it concludes the what are the Challenges for machine learning algorithms for big data analysis.

VIII. References

[1]https://www.quora.com/What-is-difference-between-Big-Data-and-Machine-Learning

[2]https://www.quora.com/What-is-difference-between-data-science-data-analytics-data-analysis

[3]https://www.quora.com/What-is-BIG-DATA-DATA-MINING-and-DATA-ANALYTICS

[4] Boyd S, Parikh N, Chu E, Peleato B & Eckstein J, "Distributed optimization and statistical learning via the alternating direction method of multipliers", Foundations Trends Mach Learn, Vol.3, No.1, (2011).

[5] Christ PF, Elshaer MEA, Ettlinger F, Tatavarty S, Bickel M, Bilic P, Rempfler M, Armbruster M, Hofmann F, DAnastasi M &Sommer WH, "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields", International Conference on Medical Image Computing and Computer Assisted Intervention, (2016), pp.415– 423.

[6] Grobelnik M, Big Data Tutorial. European Data Forum, (2013).

[7] Chen M, Xu ZE, Weinberger KQ &Sha F, "Marginalized denoisingautoencoders for domain adaptation", Proceeding of the 29th International Conference in Machine Learning, Eding burgh, Scotland, (2012).