International Journal of Computer Applications (0975 - 8887)
Volume 180 - No.20, March 2018

Big Data, Machine Learning and the BlockChain Technology: An Overview

Francisca Adoma Acheampong
School of Computer Science and Engineering
University of Electronic Science
and Technology of China

ABSTRACT

The importance of big data in machine learning cannot be overemphasized in recent times. Through the evolution of big data, most scientific technologies that relied heavily on enormous data in solving complex issues in human lives gained grounds; machine learning is an instance of these technologies. Various machine learning models that yield groundbreaking throughputs with high efficiency rates in predicting, detecting, classifying, discovering and acquiring in-depth knowledge about events that would otherwise be very difficult to ascertain have been made possible due to big data. Although big data has undoubtedly helped in the field of machine learning research ,over the years, its mode of acquisition has posed great challenge in industries,education and other agencies that obtained them for various purposes. This is because these large quantities of data cannot be stored on personal computers with limited storage capabicity but required the use of high storage capacity servers for effective storage. These servers may be owned by a group of companies or individuals who had the singular priviledge to modify the data in their possession as and when deemed relevant thus the creation of a centralized data storage environment. These were mostly refered to as the Third Parties (TP) in the data acquisition process. For the services they rendered, these trusted parties priced data in their possession expensively. The adverse effect is a limitation on various researches that could help solve a number of problems in human lives. It is worth mentioning that the security of these data being purchased expensively cannot be even assured limiting various researches that thrive on secured data. In order to curb these occurrences and have better machine learning models, the incorporation of Blockchain Technology databases into machine learning. This paper discusses the concept of big data, Machine Learning and Blockchains. It further discusses how Big data has impacted the Machine learning Community, the significance of Machine Learning and how the BlockChain Technology could be used similarly impact the Machine Learning Community. The aim of this paper is to encourge further research in incoporating the BlockChain Technology into Machine Learning.

Keywords

Big Data, Machine Learning, Blockchains, Data Preprocessing

1. INTRODUCTION

Data can be defined as a collection of values of a specific variable either qualitative or quantitative[16]. Whereas quantitative data highlights on quantity and numbers, qualitative data is more categorical and may be represented by categories such as height, color, race, gender, etc. Data is a very important resource in every research work. The type of data acquired coupled with the preprocessing techniques used contribute massively to great research achievements. Generally obtained through primary and secondary sources , data is primarily obtained by direct observations and through the conduction of surveys. Secondarily, data can also be acquired through rigorous market studies or information generated electronically or obtained from the worldwide web. Over the years, primary sources of data have provided fixed and relatively small quantities of data as compared to its secondary sources counterpart. In recent times, the acquisition of data for research projects has been made easy with the worldwide web. The massive amounts of data being generated per second through various social media platforms, online marketing platforms, and business websites among others generally defines Big Data (BD)[21]. These data may be preprocessed and analyzed upon acquisition to make better event predictions and knowledge discoveries for the benefit of man. They may also be fed into a machine learning model for automated series of specific actions. The works of R. Swathi and R. Seshadri, in [17] confirms that a solid relationship exist between machine learning and big data. This relationship is thus established from the fact that machine learning models perform comparatively better with big data than with fewer sets of data. The bigger the data, the better the classification rate, efficiency rate, prediction rate and general system throughput. Solving problems which would have been rather impossible to deal with [12, 15], Machine learning has impacted greatly in health, industry, transportation, marketing and other sectors of human lives through the development of robots to handle activities which are toxic or dangerous to humans, the timely detection of diseases such as cancer, glaucoma etc., the visualization of smart cars, effective web search, language translations and etc. Over time, the ever-increasing amount of data from different sources could not be stored on personal computers due to huge storage capacity needed and required millions of servers for appropriate storage These servers could only be owned by particular groups of companies or individuals who could afford for both their purchase and maintenance. These groups also called Trusted Parties, are trusted with voluminous amounts of data, have propri-

etary data access and release data out to individuals at a fee. BD being used to undertake machine learning projects are mostly acquired from these Trusted parties operating under centralized environments. The rippling effect is a crippling world of inventions as the purchase of data greatly limits the number and quality of research per year. Also the centralized approach greatly limits the reliability of such data because of the singular point of failure associated. In machine learning however, unreliable data means lower system throughput hence the need for much reliable data. The block chain technology may provide reliable data for machine learning projects at no charge, through a decentralized access controls approach [13]. A number of nodes are connected to each other in a form of a chain and decision making depends equally on all connected nodes i.e. No one node takes decision for the number of nodes involved hence no single point of failure [5]. The technology encourages the sharing of data between nodes. Sharing of data between nodes further imply a significantly greater amount of data within the chain. Such data can then be fed into Machine learning models directly and freely without the assistance of a trusted party that would otherwise require expensive amount of money i.e. Block chains Databases in Machine learning Models saves money. In Machine Learning, the bigger the data, the better the accuracy and greater the generalization ability of the model. i.e. Block chain implementation not only help save money but also helps in ensuring better machine learning models due to its decentralization ability cite100. In the next section, the concept of Big Data is broadly discussed, Section 3 discusses Machine learning and its associated technologies. In Section 4, the Block chain Technology is discussed showing how well it could be incoporated into Machine Learning. This paper concludes in section 5

2. BIG DATA

Big data can be defined as voluminous amount of data either structured, slightly-structured or unstructured obtained from multiple or a single source [21]. Big data is very important in making constructive research inferences, conclusions and generalizations . Most importantly big data can be efficiently mined to discover hidden patterns and obtain deeper knowledge about events. Big data is popularly characterized by the 4Vs i.e. Volume, Variety, Velocity and Veracity[18].

- **Volume:** . Large amount of data is obtained daily from the health, business, transport, entertainment as well as other important aspects of our daily living. The size of data determines whether or not it is big data.
- **Variety:** Data is being generated from different sources at the speed of light. These data from different sources are obviously of different types. The varying types of data being produced within a twinkle of an eye from different sources defines the Variety properties of big data
- **Velocity:** In the past, researchers struggled to obtain data for their work. However, with the current advancements in technology, data is being generated these days at such an alarming rate through advertising sites, marketing sites, social media platforms, and business websites among others. The rate of fast increase of data is what we characterize as the Velocity of Data. This feature has helped researchers immensely considering the fact that data acquisition now isnt as tedious as it used to be some years ago [1].
- **Veracity:** This characteristics describes how quality data should be. Making analysis with quality data goes a long way into drawing accurate conclusions.

It is worth mentioning that, big data whether structured, slightly structured or unstructured needs to be pre-processed when obtained. This helps to remove unclean, irrelevant, redundant and noisy data from the acquired data [22]. In order to obtain accurate results for a particular system/ model, data must be preprocessed.

2.1 SOME IMPORTANT ALGORITHMS FOR PREPROCESSING BIG DATA

When data is initially acquired, they may be mostly unclean, noisy, incomplete or even redundant [7]. Feeding such data into a machine learning model will produce less accurate results even with the most powerful machine learning algorithms. Hence the need for Data Preprocessing[8, 22]. Preprocessed data, coupled with appropriate machine learning algorithms produce models with high throughput and efficiency rates. Brodley and Fried in [3] placed significant emphasis on data pre-processing by showing the quality and efficient performances of models that were implemented using preprocessed data as against systems that used raw data without preprocessing. The ultimate aim of Data preprocessing is to Clean , Extract Features Data and Normalize Data.

(1) **Clean Data**: This involves removing noisy and missing or incomplete data from the acquired data.
- **Removal of Noisy Data**: Brodley and Fried [3] emphasized the importance of noise reduction by using the Ensembler Filter. Their results proved that filtering noise out of data maintained a good performance accuracy. Other Prominent algorithms for filtering noise out of data is the Iterative Partitioning Filter (IPF) proposed in [9] and the application of Denoising Autoencoders.
- **Missing Data/ Incomplete Data** : Missing or incomplete data results in inconsistencies and affect the overall performance of a system. Data may often have missing values because of unforeseen events such as incomplete downloads or failure of data collection equipments. Dealing with missing data may involve the complete removal of such data from the whole, finding statistical relationships such as the mean, median, mode etc for quantitative data and the application of other methods such as the Bayesian or Decision trees in generating new data to fill up the vacancy[6].

(2) **Extracting Features from Data**: This allows for the selection of special features from whole data ie the selection of a subset of great interest from the whole data. Through feature selection, the curse of dimensionality as a result of big data is revoked. Feature extraction helps reduce the dimensionality of data which may go a long way into increasing response time and reducing system complexities. Algorithms that facilitate feature extractions include: Principal Component Analysis, Autoencoders, Thresholding in image data, Hough Transforms etc.

(3) **Normalizing Data**: Data normalization involves organizing data in such a way as to achieve cohesion in data entities. This helps remove redundancies in data and reduce data size as well. Data after being preprocesses can then be fed into a machine learning model to perform a particular automated tasks through continuous learning.

3. MACHINE LEARNINGE

Machine Learning is an aspect of computer science that enables computers to perform specific task by learning. Through learning,

systems are able to adapt from previous experience and to perform similar or related tasks without being programmed explicitly for those tasks. Machine learning makes use of data and various algorithms in order to achieve the learning process. Some machine learning algorithms include Artificial Neural Networks, Support Vector Machines, and Nave Bayes etc. Machine learning algorithms require a reasonable amount of data in order to produce a more generalized and accurate conclusion or results [17]. Hence the link between big data and machine learning. The learning processes involved in machine learning can be supervised, unsupervised, reinforcement [19]

In Supervised Learning also called Example Learning, a models desired output is already known. It is only presented with an input example and supposed to learn to produce the intended output [11, 10]. Through various cost functions such as the cross entropy ,Quadratic and Exponential Cost, the difference between the output and intended output is found and an optimizer function such as the Adams Optimizer, Stochastic Gradient Descent(SGD) etc used to minimize such cost. Supervised learning is most often used in applications where future predictions rely heavily on historical data. For instance in predicting earthquakes.

In Unsupervised Learning, systems are expected to learn rightly from given inputs; no labels or examples are given. The system is supposed to explore very well the input data, identify patterns within and produce an output of some sort. This learning process works well on transactional data. For instance, in recommender systems.

Reinforcement Learning is commonly used in game applications where rewards or punishments are given an agent based on their actions . Agents are therefore expected to take actions to maximize their rewards by following the best policy. Reinforcement learning is composed of 3 important features. These include an Agent , Actions and the Environment. The agent is expected to performs tasks by taking actions based on their surrounding evironments. Depending on actions taken, they receive rewards or get punished. It is therefore the responsibility of the agent to apply best policies so as to increase their rewards .

3.1 Significance of Machine Learning

Machine learning has improved the quality of lives of humans by providing a number of application to facilitate human living. Among the numerous applications of machine learning in the field of health, science, industries etc. is the timely detection of diseases such as cancer, glaucoma and other diseases which are claiming human lives at a jaw -breaking rate, the visualization of smart cars, effective web search which has made the internet searches more easy, language translations are immensely helping in worldwide communications and limiting the great language barrier among countries, realization of fraud detection and face recognition systems to mention but a few are greatly helping to improve the quality of life of humans. It is in this regard that Machine Learning has remained significant over the years.

4. BLOCKCHAIN TECHNOLOGY

Blockchain is the interconnection of decentralized blocks of information [13]. The technology thrives on peer to peer networks in order to achieve its decentralization ability. In Blockchains, entries are written into a record by each peer. A number of record of information from a particular peer form a block. Each peer within the network has their own block. These blocks are interconnected to form a chain of blocks containing information[20].Information flows freely within these chained blocks.However, entries written into a record by each peer within the network of users has to be consented to by group[5]. In Blockchain technology, information is made readily available to all peers within a group or network. They then use specific protocols to determine whether an information amendment or update should or not occur. The technology derives its strength from 3 other technologies. They are Peer to Peer Network, Public Key Cryptography and the Blockchain Protocol [2].

Peer to Peer Network: Peer to Peer Technology drives the authorization and decentralization ability of the Blockchain Technology. Peers reach a consensus and decide on particular data updates or amendments. No one peer can effect change to an information without the approval of others [4].

Public Key Cryptography(PuKC): The involvement of PuKC in the blockchain technology ensures a secure digital identity. Using the associated private and public keys, a digital signature depicting strong sense of ownership could be created and hence a secure digital identity. In Public Key cryptography, a user that wishes to communicate sends a message along with its public key to a peer. The resceiving peer receives the message and uses their private key to decrypt and retrieve the message[20, 14]. This form of securing information provides high authentication access. A feature embedded in Blockchain. The authorization and authentication process involved in Block chain makes it a force to reckon with in recent times.

Blockchain Protocol:This protocol determines the underlying rules within which blockchain operates i.e. broadcasting a digitally signed information to all nodes/peers in a network at a given time. The nodes involved agree on the information update and each node/block gets a copy of the updated information hence no single point of failure. The major property of blockchain ensuring security and overall effectiveness of the technology lies with decentralization /shared controls [20]

5. BLOCKCHAINS IN MACHINE LEARNING

In order to generate good models in Machine learning, large amount of data is required. This is because large data increases the overall throughput, helps is making a more generalized conclusion and produces a more efficient and reliable system. This is one of the reasons why the importance of big data in machine learning cannot be overemphasized. However, incorporating Blockchain databases in Machine learning means having a shared data, having relatively much bigger and safer data and having much better machine learning models[2].

(1) **Shared Data**: The decentralized property of blockchains enable for data to be shared among a community of nodes. This provides easy access to data for related machine learning models implementation. The issue of data acquisition has been a major stumbling block to most machine learning researches. Previously researchers went through tough struggles to get some fixed amount of data for their research. This difficulty did not only result in the generation of less reliable and inefficient models, but also served as a major hindrance to a number of researches. With the introduction of big data, this hurdle could be crossed, however, a trusted party would be involved to get sufficiently large amount of data. These trustees would in turn be paid expensively for the data being collected. Blockchain databases however would provide data to researchers for major research projects without the services of a trusted party because of its decentralized data sharing ability. [2, 13]

(2) **Bigger and Safer Data**: Decentralized data means much bigger and safer data with data coming from both intrinsic and extrinsic sources. Intrinsic sources of data be grouped into local and metropolitan. The data that emanates from a particular place say a particular branch of a company can be said to be local. Combined Data from the same company but different branches can be termed Metropolitan data. With Blockchain , these data can be shared across and when used as input to a machine learning model, produce high efficiency rate as compared to using only locally acquired data. Extrinsic data may be data from related companies being shared. Such data when used in major predictive machine learning models can in no doubt make better predictions. Aside from acquiring voluminous amount of data through such technology at practically no expense, the data acquired is also as safe as heaven[2]

(3) **Better Machine Learning Models**: The rippling effect of getting large amount of safe data for machine learning researches is the development of better and more reliable machine learning models for various purposes as prediction, forecasting, diseases detection, voice and speech recognition, face detection, to mention but a few. [2]

6. CONCLUSION

The paper summarises briefly big data, machine learning and blockchain technology. The relevance of these technologies and how closely they relate with one another is further discussed citing major applications which makes use of these technologies together.The aim of this paper is to encourge further research in incoporating BlockChain Technology into Machine Learning.

7. REFERENCES

[1] S. Athmaja, M. Hanumanthappa, and V. Kavitha. A survey of machine learning algorithms for big data analytics. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–4, March 2017.

[2] Nolan Bauerle. How does blockchain technology work? Available at:[url = https://www.coindesk.com/information/how-does-blockchain-technology-work/,, 2018. Accessed Feb 2018].

[3] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

[4] C. Cachin. Blockchains and consensus protocols: Snake oil warning. In *2017 13th European Dependable Computing Conference (EDCC)*, pages 1–2, Sept 2017.

[5] Michael Crosby, Pradan Pattanayak, Sanjeev Verma, and Vignesh Kalyanaraman. Blockchain technology: Beyond bitcoin. *Applied Innovation*, 2:6–10, 2016.

[6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[7] S. Gharatkar, A. Ingle, T. Naik, and A. Save. Review preprocessing using data cleaning and stemming technique. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–4, March 2017.

[8] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[9] T. M. Khoshgoftaar and P. J Rebours. Improving software quality prediction by noise filtering techniques. *Comput Sci Technol*, 22:387, 2007.

[10] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[11] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[12] David J Lary, Amir H Alavi, Amir H Gandomi, and Annette L Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, 2016.

[13] W. Meng, E. Tischhauser, Q. Wang, Y. Wang, and J. Han. When intrusion detection meets blockchain technology: A review. *IEEE Access*, PP(99):1–1, 2018.

[14] James Nechvatal. Public-key cryptography. Technical report, NATIONAL COMPUTER SYSTEMS LAB GAITHERSBURG MD, 1991.

[15] M. Ngxande, J. R. Tapamo, and M. Burke. Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, pages 156–161, Nov 2017.

[16] Rod Pierce. What is data? Math Is Fun, Available at:[url = http://www.mathsisfun.com/data/data.html,, 2017. Accessed Feb 2018].

[17] A. Rathor and M. Gyanchandani. A review at machine learning algorithms targeting big data challenges. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 1–7, Dec 2017.

[18] S. R. Suthar, V. K. Dabhi, and H. B. Prajapati. Machine learning techniques in hadoop environment: A survey. In *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pages 1–8, April 2017.

[19] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[20] Karl Wst and Arthur Gervais. Do you need a blockchain? Cryptology ePrint Archive, Report 2017/375, 2017. `https://eprint.iacr.org/2017/375`.

[21] X. Wu, X. Zhu, G. Q. Wu, and W. Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107, Jan 2014.

[22] Li Xiang-wei and Qi Yian-fang. A data preprocessing algorithm for classification model based on rough sets. *Physics Procedia*, 25:2025–2029, 2012.