Big Data Analytics using Machine Learning Techniques

Shweta Mittal
Dept. of Computer Science and Engineering
Guru Jambeshwar University of Science and Technology
Hisar, India
shwetamittal019@gmail.com

Om Prakash Sangwan
Dept. of Computer Science and Engineering
Guru Jambeshwar University of Science and Technology
Hisar, India
sangwan0863@gmail.com

***Abstract*— Gigabytes of data is being generated now a days on daily basis which may possess some of the characteristics such as high speed, huge volume, uncertainty, non-stationary data, real time data etc. Conventional Machine Learning Techniques cannot be used for analysis of big data due to its above mentioned features. Also, traditional storage and processing techniques fails to meet the requirements. In this paper, we have discussed the various challenges that may occur while using traditional MLT for Big Data Analytics and its possible solutions. As per our survey, Parallel Processing, Dimensionality Reduction techniques, GPUs, Map reduce jobs, Deep learning, Online learning, Incremental learning are some of the possible solution to meet the challenges associated with big data analytics.**

Keywords—Big Data, Big Data Analytics, Machine Learning Techniques, Deep Learning

I. INTRODUCTION

With the advent of technology, data generated digitally is also increasing at a very high speed. It is not only huge in size, but is also composed of several data types, generated at a very high speed and may also contain missing or uncertain values. Any data can be classified as big data if it satisfies 5V model i.e. Volume, Variety, Velocity, Veracity and Value [1]. Big Data can be generated from various domains like Social Networking sites, Hospitality data, Meteorological data, Online Shopping sites, Banking etc. Huge data doesn't always leads to good quality results unless it is analysed effectively and efficiently. Big Data Analytics is a technique of analyzing big data sets to gain some meaningful insights so that it can be used for various business applications or in enhancing day-to-day human life.

Machine learning is the subfield of Artificial Intelligence that has capacity to learn and get better from its experiences. It learns from examples and then produces the code to perform the tasks which is in contrast with conventional way of programming. ML is being widely used now-a-days in the number of applications. Movie recommendation in Netflix, People Tagging in Facebook, Amazon's Alexa and Spam Filtering in Gmail are commonly used applications which use ML algorithm. ML techniques can be categorized as follows i.e. Supervised Learning, Unsupervised learning and Reinforcement Learning.

Supervised learning maps input parameters to its associated output with the help of certain numerical parameters. Classification and Regression are example of algorithms which use supervised learning. For predicting discrete valued attribute, Classification is used while for continuous valued attribute, Regression is used. Decision trees, Support Vector Machines (SVM), Naive Bayes algorithm etc. are examples of Classification algorithms. Linear Regression and Lasso Regression are commonly used Regression algorithms. Biometric attendance, Weather prediction, Document Classifier etc. are applications that use supervised learning technique.

Unsupervised learning learns pattern from the given dataset. Clustering, Association, Dimensionality reduction are examples of unsupervised learning. k-means, k-mediod, fuzzy means, CLARA, DBSCAN and OPTICS are commonly used clustering techniques. ANN is a technique which is common to both supervised and unsupervised learning but didn't gain attention earlier as the neurons may get trapped in local optima and hence results in premature convergence. Amazon Product recommendation is a very popular application that uses unsupervised learning algorithm.

Reinforcement learning learns by exploring its search environment and woks on hit and trial method. Fig. 1 describes commonly used machine learning techniques now a days. There are some advanced machine learning algorithm i.e. Deep Learning, Online learning, Incremental Learning, Representation learning etc which are gaining lot of popularity these days and very advantageous for big data analytics.

Deep Learning is a MLT which is inspired by the Artificial Neural Network. Deep learning provides the hierarchical representation for the data in which high level features can be represented in terms of low level features. It can be used in both supervised and unsupervised learning and has ability to learn from unlabelled datasets. A network can be considered as deep if the number of hidden layers is greater than 1. Instead of adding a new layer, a network can be trained by adding more number of hidden units, but it will not be much beneficial as network will learn only linear mapping between the variables. In 2006, Hinton *et al.* [2] proposed 2 step strategy to train neural networks leading to the popularity of Deep learning techniques. In step 1, greedy layer wise unsupervised learning is applied while in step 2, whole network is fine tuned by applying supervised learning. As number of hidden layers in the network increases, the gradient sometimes becomes too small that learning of the algorithm stops. It is known as Vanishing Gradient Problem (VGD) and may arise while training artificial neural networks [3]. To overcome VGD, LSTM (Long Short Term Memory) and faster CPU's can be used.

Deep Learning Architectures are as follows: Convolutional Neural Network (CNN), Feedfoward Network (FNN), Auto Encoders(AE) and Recurrent Neural Networks(RNN). CNN is commonly used for processing image data and comprises the following layers: Convolutional Layer, Pooling Layer and Fully Connected Layer. Network first performs several convolution operators parallely to produce set of linear activations.

978-1-5386-5933-5/19/$31.00 ©2019 IEEE

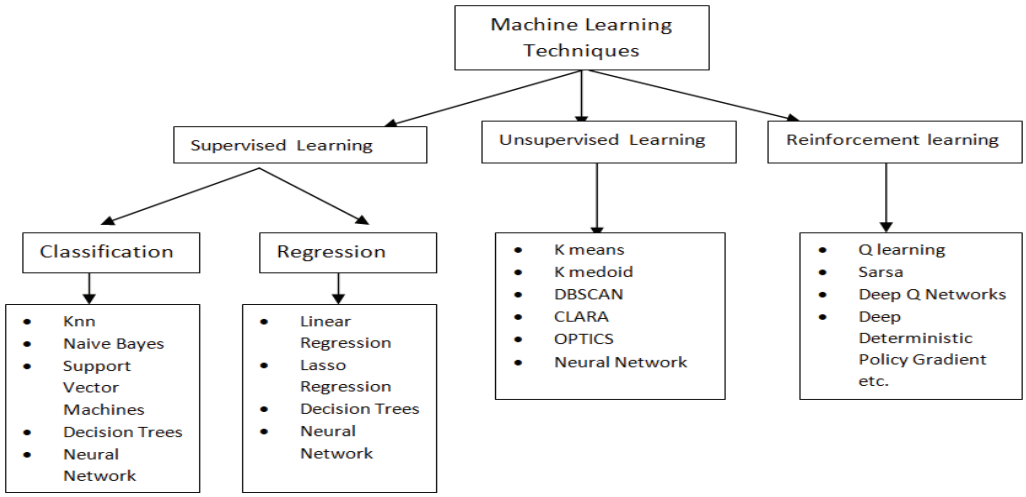

Fig. 1. Classification of Machine Learning Techniques

Afterwards, Pooling layer is applied to modify the output of layer which helps representation becomes invariant to small change in input. Fully Connected Layer is used for classification and recognition purpose on the topmost layer. Auto Encoders is unsupervised learning algorithm which attempt to copy its input to the output approximately but not exactly and has two parts: encoder and decoder. Autoencoders can be stacked to have more than 1 hidden layer. Stacked Auto Encoders, Sparse Auto Encoders, Noise Auto Encoders are some of its popular variants. Recurrent Neural Network is used to process sequential data. Key difference between FNN and RNN is that FNN do not store any previous input information while RNN takes into account the previous data. RNN may include cycles in its graph.

Restricted Boltzmann Machine (RBM), Deep Belief Networks (DBN), (DBM), Deep Stacking Networks (DSN) are Deep learning models. RBM is undirected probabilistic model which comprises of 2 layers i.e. 1 layer of hidden variables and 1 layer of visible variables. RBM has no intra layer connection and it may be stacked to form deeper layers. Deep Belief Network is first successfully trained non-convolutional model which is formed by stacking several RBM. DBN is composed of several layers of hidden variable and 1 layer of visible variable. Like RBM, it also has no intra layer connections. Deep Boltzmann Machine is composed of several layers of hidden variable and is widely used in Document Modeling. Within each layer, each of its variables are mutually independent. DSN consists of stacks of multiple modules where each module is a specialized NN.

Section 2 describes the research done by the numerous authors on ML algorithms for big data analytics. Conclusion and Future research work has been described in the Section3. Section 4 presents the scope and relevance of the review.

II. LITERATURE REVIEW

Alexandra L'Heureux *et al.* [4] presented several issues and their solutions to use MLT (Machine Learning Techniques) for Big Data Analytics. Several approaches with which machine learning can deal with big data are as follows: a) Data Manipulation which can be achieved via Dimensionality Reduction (PCA, AE, GA etc.), Instance selection and Data Cleaning, b) Processing Manipulation: By i)Vertical scaling i.e. Use of Multi-core CPUs, Supercomputers, GPU etc. and ii) Horizontal scaling: by increasing the number of nodes (use of Hadoop, Spark etc.) for processing the data. c) Algorithm Manipulation: either modify an algorithm or make existing algorithms run on parallel platforms.

Chun Wei Tsai *et al.* [5] conducted a survey and concluded that parallel computing technologies i.e. CUDA, GPU, Data reduction techniques like Sampling, Metaheuristic techniques needs to be explored for efficient big data analytics. Omar Y. Al-Jarrah *et al.*[6] presented an overview on following machine learning approaches: Ensemble learning, Local learning and Deep learning. These algorithms require least memory space and processing which makes them suitable for big data.

S. Athmaja *et al.* [7] conducted a survey of various machine learning algorithms for Big Data Analytics and concluded that Deep learning (because of its high performance), Distributed learning, Transfer learning (to process real world data), Active learning (to learn from unlabelled data), Kernel learning etc. can be used to explore various V's of big data.

Junfei Qiu *et al.* [8] also conducted a survey on big data analytics. Conventional ML algorithms are not suitable for big data analytics as they are not designed for high volume data and have the assumption that all the data is initially in the main memory. Various learning techniques like Representation learning, Deep learning, Distributed and Parallel learning, Transfer learning needs to be explored for big data. To deal with heterogeneous (i.e for different types of data) , high dimensional and non linear data, Transfer learning, Feature Reduction Techniques i.e. PCA, LDA, Representation learning and Kernel based learning can be

used. To cop up with high speed data, Online learning is a solution. For missing or ambiguous values, Advanced Deep Learning approach can be implemented. Distributed framework with Parallel computing, Parallel Programming models and Cloud computing based approaches can also be used for large scale data

Qingchen Zhang *et al.* [9] surveyed deep learning techniques for big data. As per the review, various challenges for big data analytics and its solutions has been described in the Table 1.

TABLE I. BIG DATA ISSUES AND ITS SOLUTIONS

Big Data	Models to be used
Volume	Parallel Deep learning models: DSN, DistBelief , GPU
Variety	Multimodal deep learning models, Tensor Deep learning AE
Velocity	Incremental learning, Online learning, Incremental Back propagation learning
Veracity	Denoising AE, Non local AE, Deep Imputation AE

Xue-Wen Chen and Xiaotong Lin [10] presented an overview on deep learning architectures and discussed several challenges related to several V's of big data. For massive amount of data, DSN, GPU, Large scale DBN, Large scale CNN, combination of data and model parallel schemes, COTS HPC systems has been presented.

Maryan M Najafabadi *et al.* [11] reviewed deep learning applications and its challenges in the domain of big data analytics. DL algorithms can be used for processing audio, vedio as well as sequential data and also has potential to find local and global relationship between various attributes. It has wide application in the field of Semantic Indexing, Vector Representation, Document Representation and Discriminative Tasks etc. Incremental learning, Denoising AE (DAE) , Adaptive DBN, Marginalised Stacked DAE (m-SDAE), CNN, DistBelief, GPU etc. can be used to resolve various challenges associated with big data.

Yisheng Lv *et al.* [12] implemented Stacked Auto Encoder (SAE) for traffic flow prediction taking into account historical data and real time data for three months of 2013. Performance of SAE was compared with Back Propagation Neural Network (BPNN), Random Walk (RW), SVM and Radial Basis Function NN (RBFNN) and it was concluded that SAE performs better in terms of MAE, MRE and RMSE.

X.Z. Whang *et al.* [13] implemented Multilevel Deep Learning Algorithm with co-variance analysis for finding out relationship between various sources and Adaptive Incremental Kernel Learning was used for regression. From experimental results, it was concluded that proposed DL algorithm performs better than SVM without Dimensionality Reduction.

Nagwa M. Elaraby *et al.* [14] studied DL algorithm for big data analytics. 2 types of supervised learning algorithm for deep networks are as follows: AE (Auto-Encoders) and RBM (Restricted Boltzmann Machines). Deep Belief Network (DBN), Deep Boltzmann Machines (DBM) and Deep Stacking Networks (DSN) are widely used Deep Neural Networks. The paper also highlighted the several challenges that may occur while processing the big data. From the review, it was found that it is difficult to handle real-time non-stationary data and multimodal data with the help of deep learning algorithms. Also, it is difficult to achieve parallelism as several deep learning algorithms are designed for single processor only. Storage space may be another issue for single processor systems and hence may not be suitable for big data.

Mohammad Abu Alsheikh *et al.* [15] presented an overview for handling Mobile Big data with Deep Learning techniques. Its objective was to map various activities of user like jogging, sitting, walking etc. to the accelerometer sensor available in the mobile phones. In an experimental setup, around 2.9 million labelled data and 38.2 million unlabeled data samples were used. Multilayer Perceptron, Instance based learning, Random Forests and Deep learning algorithms were implemented using Apache Spark platform. The results proved that Deep Learning has lowest error rate while Multilayer perceptron has the highest error rate. Underfitting can be avoided by using more deep networks and by increasing the number of neurons per layer. Also, with the increase in number of cores, learning time decreases.

B.M. Wilamowski *et al.* [16] presented a new visualization techniques i.e. t-SNE for visualising high dimensional data and a new clustering approach i.e. FSFDP (First Search and Find Density Patterns) has been presented. From the experimental results, it was found that FSFDP is efficient but too time consuming. P. Suresh Kumar and S. Pranaviv [17] implemented various machine learning algorithms i.e. RF, SVM, k-NN, LDA and CART on Diabetes dataset using Big Data Analytics. The above mentioned machine learning algorithms were applied on R-Studio and the results proved that RF has highest accuracy while kNN has the least accuracy.

Philipp Moritz *et al.* [18] introduced new framework i.e. SparkNet for training deep networks which is very easy to deploy and has no parameter tuning. It can also scale well with increasing number of clusters and tolerates high latency communication

Lidong Wang and Cheryl Ann Alexander [19] discussed several challenges for implementing traditional machine learning algorithms on big data. Major findings of the paper are as follows: i) Decision trees are not appropriate for big data analytics. ii) For reasonable size database, SVM is a good choice. iii) Deep learning can deal well with volume and variety aspect of big data. iv) Hadoop's Map Reduce do not support iteration feature to efficiently iterate MapReduce procedure. v) Also, machine learning algorithms are trained using certain dataset which makes them unsuitable for another dataset. Some technology progress has been made in the field i.e. Online Feature Selection, Faceted Learning (for hierarchical data structure), Multi-task Learning etc. which helps in dealing with above mentioned challenges

Veershetty Dagade *et al.* [20] studied Big Data Weather Analytics using Hadoop platform and the results proved that Hive performs better than Pig. A.K. Pandey *et al.* [21] implemented ANFIS and FL for weather prediction on

Hadoop platform and it was concluded that ANFIS is a better choice. Weizhong Zhao *et al.* [22] implemented k means algorithm on Map Reduce platform for big datasets.

Cheng-Tao Chu *et al.* [23] proposed framework for implementing machine learning algorithms using Map-Reduce platform on multicores. Algorithms that can be expressed in summation form can be easily implemented on multicore computers. Examples of such algorithms are Linear Regression, k means, Logistic regression, Neural network, Principal Component Analysis (PCA) etc. Data is split into several subsets and subsets are allocated to Mapper. Algorithms which use step of gradient ascent cannot be implemented on multicore platform. The experimental results proved that speedup is directly proportional to the number of cores. Jesus Mailo *et al.* [24] implemented MR-Knn algorithm and the comparison has been done with the sequential version.

Preeti Gupta *et al.* [25] reviewed various machine learning techniques which are scalable for big data. Parallel versions of various ML algorithms like k-means, Density based clustering, Hierarchical clustering, Decision Trees, Fuzzy rules, Random Forest has been implemented successfully by various researchers on Hadoop/Spark platform. Sampling, Indexing, Partitioning the data effectively among nodes, Number of Map-Reduce jobs, speedup, accuracy etc. are some parameters that may be considered for evaluating the performance of MLT. Summary of Machine Learning Techniques implemented by various authors and its results have been mentioned in table II.

TABLE II. MACHINE LEARNING TECHNIQUES AND ITS RESULTS

Ref. No.	Machine Learning Technique	Results
[12]	BPNN, RW, SVM and RBF	SAE performs better than BPNN, RW, SVM and RBF in terms of MAE, MRE and RMSE.
[13]	Multilevel Deep Learning Algorithm and SVM	Proposed algorithm performs better than SVM without Dimensionality Reduction.
[15]	Multilayer Perceptron, Instance based learning, Random Forests and Deep learning algorithms	Deep Learning has lowest error rate while Multilayer perceptron has the highest error rate.
[17]	RF, SVM, k-NN, LDA and CART	RF has highest accuracy while kNN has the least accuracy.
[16]	FSFDP	FSFDP is efficient but too time consuming
[20]	-	Hive performs better than Pig
[21]	FL and ANFIS	ANFIS is better than FL
[22]	Parallel k-means	Proposed algorithm is efficient
[23]	Linear Regression, k-means, Logistic Regression, NN and PCA	Speedup is directly proportional to number of cores.
[24]	MR-KNN	MR-KNN has lesser execution time

III. DISCUSSION AND FUTURE WORK

As per the review done in Section II, existing ML algorithms are not suitable for analysis of Big Data. Thus, ML algorithms needs to be optimized so that they can be effectively used for big data analytics. Optimization can be achieved by various Dimensionality Reduction techniques, Hashing, Data Cleaning, Sampling etc. Also, to minimize the execution time, parallel programming techniques can be used. Hybrid machine learning algorithms may also be advantageous in case of big data and can be explored to obtain better results. Deep learning, Representation learning, Online learning, Incremental learning etc. have been implemented successfully by researchers for big data analytics and needs further study.

IV. SCOPE AND RELEVANCE

Analysing big data has its application in numerous domains i.e. agriculture, medical, weather stations, flight delays etc. In medical field, patient records can be analysed to find out the cause of a certain syndrome. Tremendous amount of open datasets/software repositories for processing big data are available online and are increasing day by day. Some of them are as follows: National Climatic Data Centre, Healthdata.gov ,data.gov.uk, Open Source Sports etc.

Python, R, Java, Scala are most popular programming languages for machine learning. Python has several packages like NumPy, Keras, Pandas, TensorFlow, SciPy etc. which ease the scientific and numerical computations. Anaconda is an open source distribution available online for package management. Anaconda also provides IDE's like Jupyter, Spyder, rstudio etc. for quick programming. Apache Hadoop, Spark etc. are the platforms for executing Map Reduce job. With PySpark tool, one can access Ressilient Distributed Database (RDD) database in Python as it has library Py4j. Number of tools are available online to implement MLT for Big Data Analytics to help the researchers implement their work

REFERENCES

[1] M. Assuncao, R. Calheiros, S. Bianchi, M. Netto and R. Buyya, "Big Data Computing and Clouds: Trends and Future Directions," Journal of Parallel and Distributed Computing, Elsevier, vol 79–80, pp. 3-15, 2015.

[2] G. E. Hinton,S. Osindero and Y.W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, pp. 1527-1554, 2006.

[3] S. Hochreiter, "The Vanising Gradient Problem during Learning Recurrent Neural Nets and Problem Solutions," International Journal of Uncertainity, Fuzziness and Knowledge-Based Systems, 1998.

[4] A. L'Heureux, K. Grolinger, H.F. Eiyamany and M. Capretz, "Machine Learning with Big Data: Challenges and Approaches," *IEEE Access*, vol 5, pp. 7776-7797, 2017.

[5] C.W. Tsai, C.F. Lai, H.C. Chao and A.V. Vasilakos, "Big data Analytics: A Survey," Journal of Big Data, Springer, 2015.

[6] O.Y. Al-Jarrah, P.d. Yoo, S. Muhaidat, G.K. Karagiannidis and K. Taha, "Efficient Machine Learning for Big Data: A Review," Journal of Big Data Research, Elsevier, 2015.

[7] S. Athmaja, M. Hanumanthappa and K. Vasantha, "A Survey of Machine Learning Algorithms for Big Data Analytics," IEEE, International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[8] J. Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng, "A Survey of Machine Learning for Big Data Processing," EURASIP Journal on Advances in Signal Processing, December 2016.

[9] Q. Zhang, L.t. Yang, Z. Chen and P. Li, " A Survey on Deep Learning for Big Data," Elsevier, Information Fusion, pp. 146-157, 2018.

[10] X. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives," IEEE Access, vol 2, pp. 514-525, May 2014.

[11] M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar , N. Seliya, R. Wald and E. Muharemagic, "Deep Learning Applications and Challenges in Big Data Analytics," Journal of Big Data, SpringerOpen, 2015.

[12] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Wang, "Traffic Flow Prediction with Big Data: A Deep Learning Approach," IEEE Transactions on Intelligent Transportation Systems, vol 16, pp 865-873, April 2015.

[13] X.Z. Wang, J. Zhou, Z.L. Huang, X.L. Bi, Z.Q. Ge and L. Li, "A Multilevel Deep Learning Method for Big Data Analysis and Emergency Management of Power System," IEEEXplore, International Conference on Big Data Analysis, 2016.

[14] N.M. Elaraby, M. Elmogy and S. Barakat, "Deep Learning: Effective Tool for Big Data Analytics," International Journal of Computer Science Engineering, vol 5, pp. 254-262, September 2016.

[15] M.A. Alsheikh, D. Niyato, S. Lin, H. Tan, and Z. Han, "Mobile Big Data Analytics using Deep Learning and Apache Spark," IEEE Network, vol 30, no. 3, pp. 22-29, 2016.

[16] B.M. Wilamowski, B. Wu and J. Korniak, "Big Data and Deep Learning, "IEEE International Conference on Intelligent Engineering Systems , pp. 11-16, 2016.

[17] P.S. Kumar and S. Pranaviv, "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics," IEEE, International Conference on Infocom Technologies and Unmanned Systems (ICTUS), 2017.

[18] P. Moritz, R. Nishihara, I. Stoica and M.I. Jordan, "Sparknet: Training Deep Networks in Spark," International Conference on Learning Representations, ICLR, 2016.

[19] L. Wang and C.A. Alexander, "Machine Learning in Big Data," International Journal of Mathematical, Engineering and Management Sciences, vol 1, no 2, pp. 52–61, 2016.

[20] V. Dagade, M. Lagali, S. Avadhani and P. Kalekar, "Big Data Weather Analytics using Hadoop," International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), April 2015.

[21] A.K. Pandey, C.P. Agrawal and M. Agrawal, "A Hadoop based Weather Prediction Model for Classification of Weather Data," IEEE, Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2017.

[22] W. Zhao, H. Ma and Q. He, "Parallel K-Means Clustering Based on MapReduce," CloudCom, Springer, 2009.

[23] C. Chu, S.K. Kim, Y. Lin, Y. Yu, G. Bradski, A.Y. Ng and K. Olukotun, "Map-Reduce for Machine Learning on Multicore," 19th International Conference on Neural Information Processing Systems, 2006.

[24] J. Maillo, I. Triguero and F. Herrera, "A MapReduce-based k-Nearest Neighbor Approach for Big Data Classification," IEEE Trustcom/BigDataSE/ISPA, 2015.

[25] P. Gupta, A. Sharma and R. Jindal, "Scalable Machine-Learning Algoritms for Big Data Analytics: A Comprehensive Review", John Wiley & Sons Ltd., vol 6, 2016.