International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-8, Issue-3, May 2020
https://doi.org/10.21276/ijircst.2020.8.3.34
www.ijircst.org

A Review of Machine Learning Techniques over Big Data Case Studies

Dr Yojna Arora

ABSTRACT- In the recent years, Data has increased exponentially and is termed as Big Data. Data Amount, Data Speed and Data Variation are three major parameters of Big Data. There are many challenges which have tuned up out of which Data Storage, Data Analysis and Data Management are the biggest ones. In order to deal with these challenges, Machine Learning, a subset of Artificial Intelligence provides various tools and techniques. This paper gives a detail about Big Data and Machine Learning. It also includes detailed literature review on various Big Data case studies which are solved by Machine Learning Techniques.

KEYWORDS- Big Data, Data Analytics, Machine Learning, Deep Neural Network, Supervised Learning, Neural Net, Data Mining, Computing

I. INTRODUCTION TO BIG DATA

A. Big Data Definition

Big Data refers to extremely large, very fast, highly diverse and complex data that cannot be managed with traditional data management tools [1],[2],[3], [4]. It is being generated across the globe at an unexpected speed. Big Data has 5 dimensions associated with it. These are Volume, Variety, Velocity, Value and Veracity. The term was evolved from 3 dimensions to 4 dimensions and now 5 dimensions define Big Data. Big Data can be examined at two levels Basic Level and Advanced Level. At the lower or basic level it is assumed as any other collection of data which can help in Business Analytics. On the other hand it is assumed as a special type of data which has great challenges and great benefits. Big Data is different from traditional data in every way i.e space, time and function. Also, Big Data is not just data in the form of rows and columns rather it includes text, audio, images, videos and other varied data representation formats. Big Data is majorly unstructured in nature.

Manuscript Received April 21, 2020

Dr. Yojna Arora, Assistant Professor, Department of Computer Science & Engineering, Amity School of Engineering & Technology, Amity University, Haryana, Manesar, Gurgaon, Haryana, India (email: yojana183@gmail.com

B. Characteristics of Big Data

Big Data is not just huge amount of data or data coming at high speed or data coming in various formats or data in doubt rather it is a combination of all. These define the V's of Big Data. Big Data was initially characterized by 3 V's [5], then 4th V was added [6] and now a 5 V Architecture of Big Data is defined [7]. There characteristics are explained as below:

i. Volume: The data is getting exponentially generated. It is almost doubling in every 12-18 months. Earlier data was measured in GBs and TBs but now the date has increased to such an extent that it is measured in Petabyte (PB) and Exabyte (EX). The data is so huge in nature that it is almost impossible to lookup some information into it in a reasonable period of time. Thus, Volume is one important factor which has made Data as Big Data

ii. Velocity: The second V of Big Data refers to Velocity. Velocity means that the data is arriving at a very high speed with no control over it. This speedy generation of data is due to easy and speedy access to internet. Data is generated from many devices and communicated very fast. Managing this highly speedy data and finding relevant information form it is a difficult task.

iii. Variety: Variety refers to data in different forms and formats. There are three aspects based on which data is categorized. These three aspects are Form of Data, Function of Data and Source of Data. **Form of Data** refers to data in different formats i.e Text, Audio, Images, Video, Graph, Map etc. or a combination of any two or more formats. Each format has its differ storage capacity and analysis complexity. **Function of Data** refers to data like Human Conversation, Songs, Transaction data, Machine Operation Data etc. All these data have to analyze in different way with different result expectations. Lastly **Source of Data** refers to data coming from different sources such as Structured, Semi Structured and Unstructured Data, Structured Data is the most organized form i.e. in the form of rows and columns, Semi Structured is partially organized data such as XML files, Log files etc., Unstructured Data is the most unorganized form which includes Text, Audio, Video, Images.

iv. Veracity: Veracity refers to data in doubt. It takes Data Quality into consideration. Since, Big Data is huge in nature so it difficult to maintain the truthfulness and quality of data reducing the noise. The reason behind depreciation of data quality may be unauthorized data source, human or machine generated errors or it can be an intentional attempt

Copyright © 2020. Innovative Research Publication. All Rights Reserve

to hamper data. This characteristic has greater emphasis because degree of data quality can ensure its applicability in various domains.

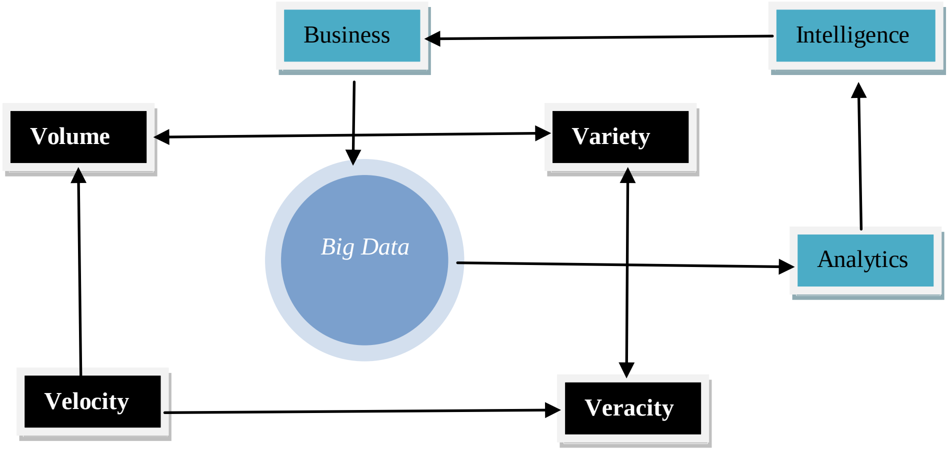

Fig 1: Big Data Environment

C. Big Data Architecture

Big Data Architecture contains three logical layers: Data Source as Input Layer, Data Processing as Middle Layer and Data Consumption for Result Analysis and Interpretation. This generalised Big Data Architecture is resistant, secure, cost effective and adaptive in nature. Major organizations have modified the architecture according to optimised infrastructure and requirements. The function of each layer of Big Data Architecture is explained as:

i. Source Layer: The source of data can be identified based on the application over which analysis is to be performed. It will greatly vary in its speed, size, form, function etc.

ii. Ingest Layer: This layer receives the data coming from various sources, in different amount and at variable speed. It the decides whether the data has to be send for Batch Processing, Stream Processing or stored in underlying database.

iii. Batch Processing Layer: It receives data from Data Ingest Layer, File System or NO SQL. It process the data using parallel processing techniques

iv. Stream Processing: It receives data only from the Ingest Layer. It works on Real Time Data which is getting continuously generated and produce desired results

v. Data Organization Layer: This layer further receives data from both Batch Processing and Stream Processing Layer. It is referred as NoSQL database. This layer is added to organize the data for easy access.

vi. Infrastructure Layer: Infrastructure Layer provides all the basic support including storage, computation and communication support.

vii. Distributed File System: This is the underlying data source which can store huge amount of data. It provides the data to all other layers.

viii. Data Consumption Layer: This layer is the final layer. It receives the output from organizing layer and provides the output in the form of reports, graphs and visualization methods.

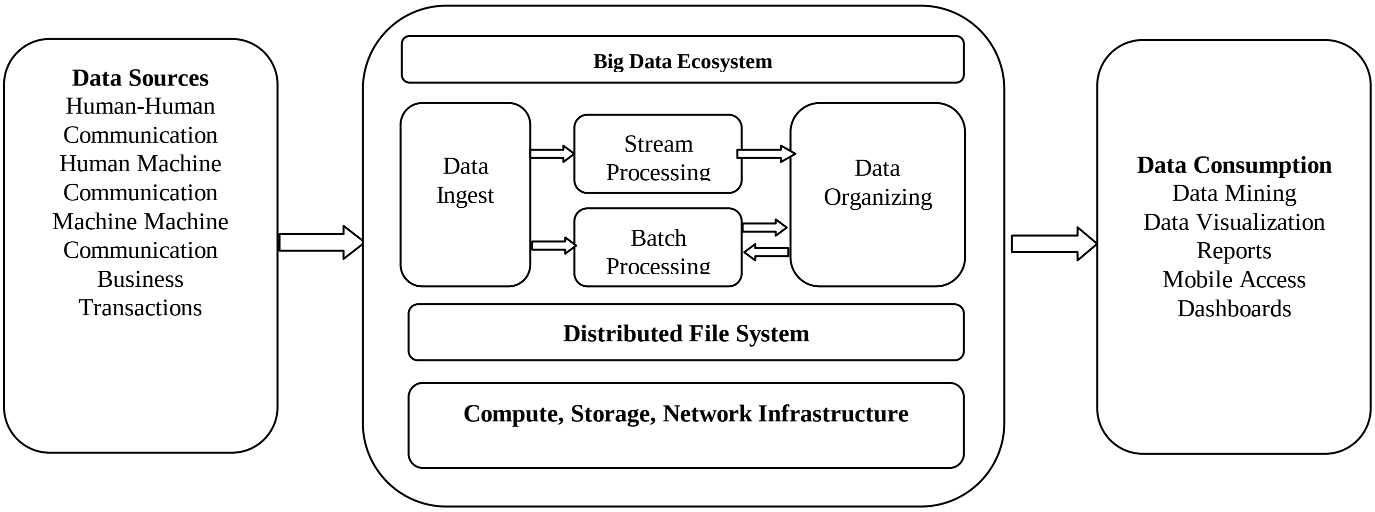

Fig 2: Big Data Architecture

II. INTRODUCTION TO MACHINE LEARNING

A. Basic Definition

Machine Learning is an application of Artificial Intelligence which allows the system to learn automatically with experience without being programmed. The learning process comes with observation of data. It helps in decision making by studying patterns in datasets. Machine Learning Algorithms can be categorized as Supervised, Unsupervised, Semi Supervised and Reinforcement [12]

Supervised Learning algorithms apply the predefined knowledge on the new set of data. It follows the Data Classification method. It labels the data set and based on it classifies the new data which come for analysis [12] Unsupervised Learning Algorithms do not follow classification or labelling approach rather it can generate inferences.Semi Supervised Learning Algorithm is a combination of both Supervised and Unsupervised Learning. It uses the approach of both labelled and unlabelled data for training Reinforcement Learning Algorithm allow the system to learn from its environment and gain knowledge

Table 1: Machine Learning Algorithms

Regression Algorithms	**Instance Based Algorithms**	**Regularization Algorithms**	**Decision Tree Algorithms [8]**	**Bayesian Algorithms**	**Clustering Algorithms [9]**	**Ensemble Algorithms**	**Artificial Neural Network [10]**	**Deep Learning Algorithms [11]**	**Dimensionality Reduction**
Ordinary Least Square Regression	K Nearest Neighbours	Ridge Regression	Classification & Regression Tree	Naïve Bayes	k- Means	Boosting	Perceptrons	Convolutional Neural Network	Principal Component Analysis
Linear Regression	Learning Vector Quantization	Least Absolute Shrinkages and Selection Operator	Iterative Dichotomiser	Gaussian Naïve Bayes	K-Medians	Bootstrapped Aggregation	Multilayer Perceptrons	Recurrent Neural Networks	Principal Component Regressio
Logistics Regression	Self Organizing Maps	Elastic Net	C4.5 and C5.0	Multinomial Naïve Bayes	Expectation Maximization	AdaBoost	Back-Propagation	Long Short-Term Memory Networks	Partial Least Squares Regression
Multivariate Adaptive Regression	Locally Weighted Learning	Least Angle Regression	Chi-squared Automatic Interaction Detection	Averaged One-Dependence Estimators	Hierarchical Clustering	Weighted Average	Stochastic Gradient Descent	Stacked Auto-Encoders	Sammon Mapping
Locally Estimated Scatterplot Smoothing	Support Vector Machines		Decision Stump	Bayesian Belief Network		Stacked Generalization	Hopfield Network	Deep Boltzmann Machine	Multidimensional Scaling
Step Wise Regression			Conditional Decision Trees	Bayesian Network		Gradient Boosting Machines	Radial Basis Function Network	Deep Belief Networks	Projection Pursuit

B. Big Data and Machine Learning Altogether

Big Data is the huge amount of data which is getting generated at a very high speed and in various formats. Analyzing this varied data is the biggest challenge. This data analysis helps in identifying hidden patterns which can help in taking better business decisions. Machine Learning on the other hand is a subset of Artificial Intelligence which helps the machine in taking future decisions based on the information which is already fed into it.

Both Big Data and Machine Learning are mutually dependent on each other. Big Data provides the datasets and Machine Learning provides methods and techniques which can be applied to analyze that data. Big Data deals with storage, ingestion and extraction tools however, machine Learning deals with prediction methods. A detailed literature review on application of Machine Learning techniques on Big Data by various researchers is shown in table below.

Copyright © 2020. Innovative Research Publication. All Rights Reserve

Table 2: Big Data & Machine Learning Case Studies

Author's Name	Aim	Technique Applied	Key Features	Advantages	Results Attatined
Yisheng Lv et al [13]	To propose a Traffic Flow Prediction Model	Deep Learning Approach with SAE Model	Use of Auto encoders as building blocks Greedy Layer wise unsupervised algorithm	Model can discover Latent traffic Flow feature representation	Proposed model performed superior than Back Propagation, Support Vector Machine and RBF Neural Network
Machine Learning fro Data Mining (Paper 3)					
Breiman, L. et al [14]	To build a Decision Tree	Recursive Partition Tree	Classification & Regression Tree	Gain of information for each feature	Decision tree used for Regression
Altman, N. S [15]	To implement algorithm for Memorizing new items	K Nearest Neighbour	Non Parametric Regression	Not strongly dependent on shape of Regression Function	Parametric models are implemented for data description
Russell S et al [16]	To compute probability of new item belonging to a class	Bayesian Networks	Bayes Theorem	Consideration of independence or dependence of features	The probability of new item is computed based on the values of features of each item belonging to each class
Cortes C et al[17]	To implement Classification Model	Support Vector Machine	Data Representation in Hyper Space	Provide good support for unknown data sets. Best suited for semi structured and unstructured data	Respective hyper planes that better divide different classes
Bishopp C.M [18]	To classify, predict or label data	Artificial Neural Network	Connecting Neurons and Connecting Layer	Ability to work with incomplete information	A self learning model is implemented to classify and predict data
Jianpeng Qi et al [19]	To implement Clustering Algorithms	K Means	Random Selection	Easy adaption to new data sets	Group Data by similarities and highlight differences and similarities between groups found
Syoji Kobashi et al [20]	To implement a postoperative prediction model	Feature Extraction using Support Vector Machine Prediction using Machine Learning	Principal Component Analysis	Helps in Pre Operative Planning	The performance of prediction model was evaluated based on correlation coefficient and root-mean-squared error
Aras Can Onal et al [21]	To implement a framework for Weather Data Analysis	Weather Clustering Sensor Anomaly Detection	Implementation of k means clustering algorithm	Integration of data retrieval, processing and learning layer	Meaningful information is extracted using the proposed framework
J. L. Berral-Garcia [22]	To study various machine learning algorithms	Decision tree algorithms, K-Nearest neighbor algorithms, Bayesian algorithms,SVM, ANN, K-means,	Execution framework and tools, platforms and libraries are explained	Detailed description about all machine learning algorithm for big data analytics	Analusis of machine learning algorithm fr Classification, Prediction and Modelling
J. Qui, Q. Wu et al [23]	To study various Machine Learning algorithms on Big Data	Gaussian Mixture models, Hidden Markov Models, SVM, logistic regression, Kernel Rgression,	Deep neural networks, Deep belief networks	Supports in identification og patterns and trends	Various traditional and new machine learning algorithms over Big Data are analyzed
M.U. Bokhari et al [24]	To propose a model for Big data Storage and analysis	HDFS for Data Storage ANN, SVM for analysis	Layered Architecture Model	Combination of Big Data Technology for Storage and Machine Learning for analysis	A 3 layer architecture model is implemented.
P. Y. Wu et al [25]	To analyze Biomedical Big	Logistic regression, PCA, HMM, Local	Case study taken from real biomedical	More accurate prediction	Big Data Analytics over biomedical data helped in

	Data	regression, cox regression	data		precision medicine
M. R. Bendre et al [26]	To implement a prediction model	Map Reduce Linear Regression	Supervised Learning Approach	Model implementation based on past records	A model is implemented for better prediction of rainfall
Ananthi Sheshasaayee et al [27]	To develop a model for temperature prediction	Apache Spark based model	Tree based machine learning algorithm for training data	Map Reduce Method of parallelizing is replaced	The proposed model optimizes the machine learning technique in a distributed environment
Junfei Qiu et al [28]	To integrate Big Data Analysis using Machine Learning method	The functionalities of Apache Spark MLib	Regression, Classisfication, Dimension Reduction and Rule extraction	Open source, Scalable, Platform independent machine learning libraray	Various qualitative an d quantitative attributes of the library are analyzed using real world data sets

III. CONCLUSION

The paper addresses the problem of Big Data and mentions its tool as Machine Learning. Initially, the paper explains about basic Big Data terms, its definitions, its basic characteristics as Volume, Variety and Velocity. It further shows basic Big Data Architecture which can be used by various organizations and modified according to their requirements. The later part of the paper explains about Machine Learning which is a subset of Artificial Intelligence. Machine Learning provides various algorithms which can be used to deal with Big Data problems. Lastly, a detailed literature on Big Data Case Studies and its respective Machine Learning technique is mentioned.

REFERENCES

[1] Stephen Kaisler, Frank Arrmour, J. Alberto," Big Data: Issues and Challenges Moving Forward",46th Hawaii International Conference on System Science, IEEE,2012

[2] Sam Padden, "From database to Big Data,", in IEEE Computer Society, 2012

[3] Dan Garlasu, "Data Implementation Based on Grid Computing",11th RoEdunet International Conference, IEEE, 2013.

[4] Avita Katal, Mohammad Wazid and R H Goudar, "Big Data: Issues, Challenges, Tools and good Practices", in IEEE 2013

[5] Doug Laney, "3 D Data Management : Controlling Data Volume, Velocity and Variety", in Application Delivery Stratergies, Meta Group, 2001

[6] First Tekiner and John A keane, "Big Data Framework", in IEEE international conference on Systems, Man and cybernetics, IEEE, 2013

[7] Parth Chandarana and M Vijayalakshmi, "Big Data Analytics Framework", in International Conference onCircuits, System, Communication and Information Technology Applications",IEEE, 2014

[8] Anuja Priyama, Abhijeeta , Rahul Guptaa , Anju Ratheeb and Saurabh Srivastavab, "Comparative Analysis of Decision Tree Classification Algorithms", International Journal of Current Engineering and Technology, Vol 3, No 2, June 2013

[9] Prof. Neha Soni & Prof. Amit Ganatra, "Categorization of Several Clustering Algorithms from Different Perspective: A Review", International Journal of Advanced Research in Computer Science and Software Engineering", Volume 2, Issue 8, August 2012

[10]Amanpreet Singh ; Narina Thakur ; Aakanksha Sharma, "A Review of Supervised Machine Learning Algorithms", 3rd International Conference on Computing for Sustainable Global Development , IEEE, 2016

[11] Ajay Shrestha & Ausif Mahmood, "Review Of Deep Learning Algorithms And Architectures", Vol 7, Ieee Access, 2019

[12] Ayon Dey, "Machine Learning Algorithms: A Review", International Journal of Computer Science and Information Technologies, Vol 7, 2017

[13] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang, Fellow, "Traffic Flow Prediction With Big Data: A Deep Learning Approach", IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 16, NO. 2, APRIL2015

[14] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. "Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. 1984.

[15]Altman, N. S. "An introduction to kernel and nearest-neighbor nonparametric regression".The American Statistician 46 (3): 175–185. 1992

[16] Russell, S.; Norvig, P, "Artificial Intelligence: A Modern Approach" (2nd edition). Prentice Hall, 2003.

[17]Cortes, C.; Vapnik, "Support-vector networks. Machine Learning" 20 (3): 273, 1995.

[18] Bishop, C.M,"Neural Networks for Pattern Recognition", Oxford: Oxford University Press. 1995.

[19]Jianpeng Qi et al, "An effective and efficient hierarchical *K*-means clustering algorithm", International Journal of Distributed Sensor Network", 2017

[20]Syoj Kobashi, Belayat Hossain, Manabu Nii, Syunichiro Kambara, Takatoshi Morooka, Makiko Okuno & Shiichi Yoshya, "Prediction of Post

Copyright © 2020. Innovative Research Publication. All Rights Reserve

Operative Implanted Knee Function using Machine Learning in Clinical Big Data, International Conference on Machine Learning and Cybernatics, 2016

[21] Aras Can Onal, Omer Berat Sezer, Murat Ozbayoglu &Erdogan Dogdu†. “Weather Data Analysis and Sensor Fault Detection Using An Extended IoT Framework with Semantics, Big Data, and Machine Learning”, International Conference on Big Data, 2017.

[22] J. L. Berral-Garcia, “A quick view on current techniques and machine learning algorithms for big data analytics”, 18th International Conf. on Transparent Optical Networks, pp.1-4, 2016

[23] J. Qui, 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, Springer, vol. 2016:67, pp. 1-16, 2016

[24] M. U. Bokhari, M. Zeyauddin and M. A. Siddiqui, “An effective model for big data analytics”, 3rd International Conference on Computing for Sustainable Global Development, pp. 3980-3982, 2016

[25] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman and M. D. Wang, “–Omic and Electronic Health Record Big Data Analytics for Precision Medicine”, IEEE Transactions on Biomedical Engineering, vol. 64, issue 2, pp. 263-273, 2017

[26] M. R. Bendre, R. C. Thool and V. R. Thool, “Big data in precision agriculture: Weather forecasting for future farming”, 1st International Conf. on Next Generation Computing Technologies, pp. 744-750, 2015.

[27] Ananthi Sheshasaayee & J V N Lakshmi, “An insight into Tree Based Machine Learning techniques for Big Data Analytics using Apache Spark”, International Conference on Intelligent Computing, Instrumentation and Control Technologies, 2017.

[28] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng, “A survey of Machine Learning for Big Data Processing”, Journal of Advances in Signal Processing, 2016

Copyright © 2020. Innovative Research Publication. All Rights Reserve