March - April 2020
ISSN: 0193-4120 Page No. 2071 - 2076

A Study on First Ordered Online Supervised Learning Techniques and Algorithms for Big Data Analytics

1. N. Sai Lohitha, Research Scholar, SCOPE, VIT University, Vellore
2. Dr. M. Pounambal, Assosciate Professor, SITE, VIT University, Vellore

Article Info
Volume 83
Page Number: 2071 - 2076
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 18 March 2020*

Abstract
Big Data Analytics and Machine Learning techniques are emerging streams both in sciences and Industry as well. The characteristics of big data include volume, velocity, visualization, viscosity and variety. Machine learning is at its crux because of its ability to analyse data and provide accurate results. With steady data growth rate, machine learning process can learn more quickly and present promising results. Due to this reason big data analytics and machine learning facilitate each other. Traditional machine learning algorithms like supervised, unsupervised, Reinforcement techniques were not so effective on big data. Map reduce, Batch learning, Online learning, Incremental learning are few of the possible solutions to fit the challenges correlated with big data analytics. The area of online machine learning in big data streams covers algorithms that work on data streams with a limited scope to store past data. In data stream model, past data does not exist to make the heuristic decisions as the fresh data is generated. In this study, an overview of machine learning models for online learning is provided. The most important ideas for classification, regression, recommendation, and supervised modeling from streaming data has been highlighted.

Keywords: *Map Reduce, Machine Learning, Supervised Learning, Big Data, Reinforcement techniques.*

I. INTRODUCTION

In the current world, the data across the globe is being exchanged at an exponential rate because of the social media, IoT sensor devices etc. For example, Facebook generates 4 petabytes of data per day—that's a million gigabytes [1]. Big Data has plays a vital role in areas like biology, transportation, online advertising, energy management and financial market[2],[3]. Mining and extracting meaningful patterns from massive input data for decisionmaking,prediction, and other deductions is the key role of Big Data Analytics. Apart from analyzing massive volumes of data, Big Data Analytics give raise to other unique challengesfor machine learning and data analysis, including format variation of the raw data, fastmoving, streaming data [4]. Unfortunately, as the volumeof data is growing, the collection of data sets is so large and complex that it is difficult to deal with using traditional learning methods since the process of learning from prominent datasets was not designed to and will not work well with bulk data. For instance, most traditional machine learning algorithms are designed for data that would be completely loaded into memory [5], which does not hold any more in the context of big data. Therefore, online learning from these numerous data is significant.

A. Machine Learning

Machine Learning is to train a machine to learn from past and predict future based on the knowledge acquired. The objective is to formulate learning algorithms that do the learning naturally without human assistance or supervision. Machine Learning is a sub area of artificial intelligence which empowers software applications to get into a state of self-learning without being programmed explicitly [6][10]. When presented to new data, these systems are empowered to learn, change, grow and implement by themselves. Machine Learning concentrates on the advancement of systems that can get to information and utilize it from themselves [7].

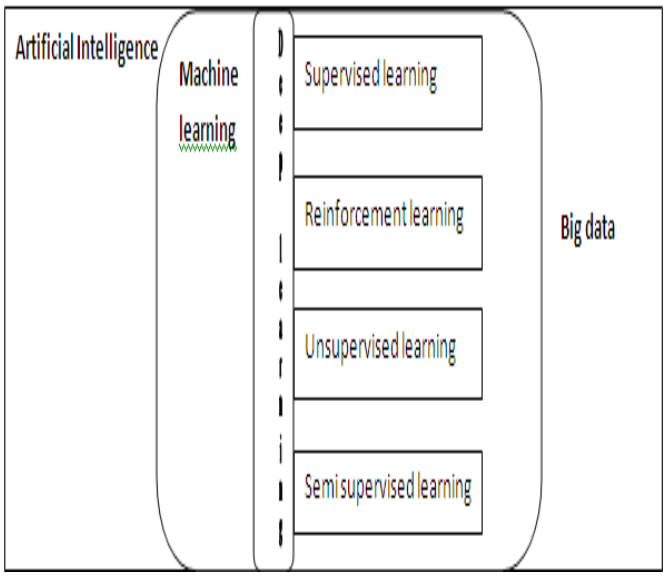
Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2071 - 2076

Machine Learning helps to identify the data and trends. Designing a machine learning application involves four steps. 1. Collect the data and prepare it. 2. Choosing the model 3. Train the data set by choosing learning algorithm 4. Evaluation and Prediction.

B. Big Data

Big Data is an emerging term which deals with enormous data that can be of structured, semi-structured and unstructured data. The analytics that include a wide range of facilities from basic data mining to advance machine learning is known as Bigdata Analytics[9]. The 4 V's of big data are Volume, Velocity, Variety and veracity. The applications are Data Mining, Business Intelligence, Visualization, Statistical Analytics, Machine Learning. The Big Data Analytics life cycle consists of 9 steps i.e 1. Business case Evaluation 2. Data Identification 3. Data Acquisition and Filtering 4.Data Extraction 5. Data Validation and Cleansing 6. Data Aggregation and Representation 7. Data Analysis 8.Data Visualization 9. Use of Analysis Results [9].



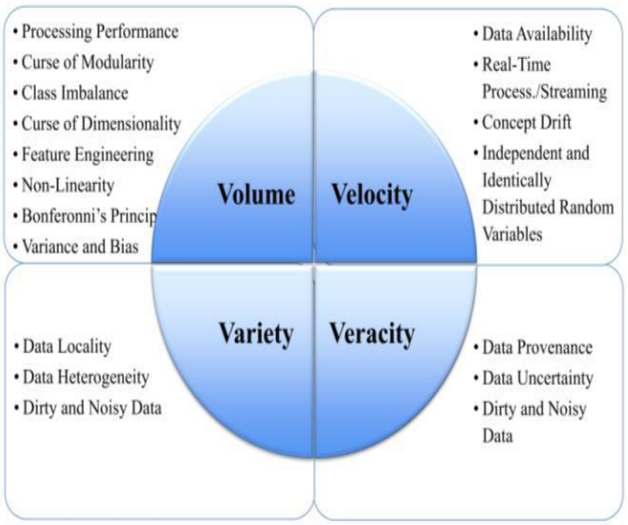


Fig 1: Machine Learning in Big Data

II. Related Work

This section widely represents the research experimentation and advancements that was done over past years. Alexandra L'Heureux, discussed the issues that are favourable relevant to Machine learning in accordance with Big Data and reviewed how emerging trends are reciprocating to these issues [11]. Junfei Qiu, assesed the current Machine Learning techniques and imported some of modern learning methods to figure out Big Data problems and summarized the challenges, research related trends, open issues of Machine Learning techniques for handling Big Data [15]. Mounica Vennapusa proposed that online learning in big data needs more amount of time for training on an isolated system. Distributed learning with a more number of systems minimizes the retained effectiveness per system and affects the whole performance [16]. Youming, proposed reference anatomy of Machine Learning for Big Data Analytics; analyzed research challenges and issues of Machine associated with Big Data [17]. Sreenivas R. Sukumar critiqued the Machine Learning challenges and interpreted these challenges of Machine Learning in Big Data era [13]. Nagwa M. Elaraby, reviewed Deep Learning architectures

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2071 - 2076

and its usage, challenges on Big Data Analytics [14]. K.Grolinger mentioned the need to combine streaming solutions with machine learning algorithms to provide instantaneous results [11]. Roheet Bhatnagar, proposed a review of vital challenges, recent advancements in Machine Learning for Big Data Analytics [19], explored how Deep Learning context and how it can be used for addressing some Key challenges in Big Data Analytics [20].

III. Anatomy of Machine Learning

The importance of real-time data in the present technological era of sensor devices, mobile devices, and social media has lead to the emergence of streaming systems which include Twitter, you tube and facebook. However, in the context of streaming data, where new data are constantly arriving, such a requirement cannot be fulfilled. Moreover, even data arriving at non-real-time intervals may pose a challenge. In machine learning, a data is typically trained by using algorithms and then performs the learned task, for example classification or regression, on new data. In this situation, the model does not automatically learn from streaming data, instead accomplish already learned task on new data. To accommodate the knowledge present in fresh data, these models should be retrained. Without re-training, the model may become outdated and doesn'treflect the recent state of the system. Therefore, to adapt to live data, algorithms must support some learning techniques which are mentioned in section IV.

3.1 Machine Learning Algorithms in Big Data Analytics

Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y).

Y = f(X) is used to make effective predictions of Y for X. The main aim is to predict the results accurately. In big data, due to huge amount of data, performing analysis on data will be possible through machine learning. With this machine learning we can extract efficient patterns.

The following brief list identifies the different Machine Learning algorithms when applied in Big Data.

1. **Decision tree** is a supervised learning method which can be usedfor classification or regression. In Decision Tree Learning, a training model is created and it determines the outcome value by learning decision rules deducted from the data attributes [21]. These decision tree algorithms have several limitations like whenever the data is very huge, creating a decision tree is time consuming. The communication cost gets increased as the data distribution is not maximized.

2. **Support Vector Machine** is a supervised learning method which can be utilized either for classification or regression [23]. When applied to big data, the SVM technique is not efficient because of its high computational complexity. For enormous amount of data, the computational requirement and storage will gets increased rapidly.

3. **K-Nearest Neighbor (KNN)** algorithm is also called as lazy algorithm. It chooses the nearest neighbours based upon the distance metric. As the number of nearest neighbors is increased, value of k, accuracy gets increased. This also effects the sensitivity and specificity.For bigdata applications, KNN is not pragmatic due to its high computation and increased memory usage cost.

4. **Naive bayes** is a strongclassifier for classification task. It identifies membership probabilities of each class or data point that comes under certain class. The class with the highest probability is evaluated as the most likely class. In Big data dueto the text redundant features and rough parameter estimation the performance of this classifier is not attainable in text classification.

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2071 - 2076

5.Neural Networks, a semi supervised technique that is used for classification and regression. Neural Networks is a computing system consists of highly interconnected processing elements which processes data by their dynamic state response to external inputs [22]. A neural network will take the input data and push them into secondary layers. For Bigdata with huge amount of data, neural networks has got few challenges. Enormous amount of data makes the technique tedious to perpetuate both reliability and effectiveness. Due to redundancy, huge workload is laid on the system.

Algorithm	Accuracy	Tolerance to missing values	Tolerance to irrelevant data	Tolerance to noise	Handling overfitting	Speed
Decision Tree	Low	Medium	Medium	Low	Low	Fast
Support Vector Machine	High	Low	High	Low	Low	Fast
Neural Networks	Medium	Low	Low	Low	Low	Fast
Naive Bayes classifier	Medium	High	Low	Medium	Medium	Fast
K-Nearest Neighbour	Low	Low	Low	Low	Medium	Low

Fig 2: Comparison of Machine Learning Techniques

IV. Advanced Learning Techniques

As the data size is increasing exponentially, the machine learning methods need to scale up. The techniques are 1. Deep Learning 2. Feature Learning 3. Active Learning 4. Ensemble Learning 5. Active Learning 6. Online Learning 7. Distributed Learning 8. Transfer Learning[16].

Online Learning

Online learning is a specialization of machine learning and includes an important family of learning techniques which are devised to learn models incrementally from data in a sequential manner[25].Online learning overcomes the limitationsof traditional batch learning where the model can be updated instantaneously and efficiently when new training data arrives. Besides, online learning algorithms arequite easy to understand, simple to implement. Due to necessity of making machine learning practical forstreaming big data analytics, online learning has become dominant in recent years. In order to react to a live data and train model over time, Machine Learning research team typically do one of these two possibilities like 1.They manually train on streaming data, and deploy the resulting model and if the prediction is as expected, they schedule training on new data to happen frequently and automatically deploy the resulting model.Online learning algorithms are more efficient and scalablefor large-scale data and also data coming with high velocity.

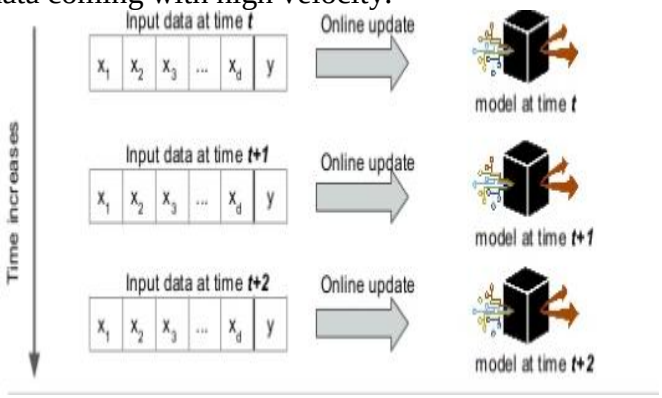


Fig 3: Online Learning Process

There are two target functions in online learning 1. Stationary Targets 2. Dynamic Targets.

4.1 Online Learning Techniques

Online learning techniques can be classified into the following three major categories:

1. Online supervised learning: In supervised learning technique full feedback information is always revealed to a learner at the end of eachonline learning round. It can be further classified into two groups of studies: (i) OnlineSupervised Learning which forms the basic knowledge on techniques and principles forOnline Supervised Learning; and (ii)Applied Online Learning" which contain more non-traditional online supervised learning techniques. Family of this learning technique include i) First-order online learning techniques ii) Second-order online learning techniques iii) Prediction with expert advice iv) Online learning with regulation [24].

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2071 - 2076

2. Online learning with limited feedback: This is concerned with tasks where anonline learner take partial feedback during theonline learning process.

3. Online unsupervised learning: In this technique online learning tasks receives sequence of data occurrence without any additionalfeedback during the online learning task. Unsupervised onlinelearning can be assesed as an add on to traditional unsupervised learningused to deal with data streams, which is typically studied in batch learning fashion.Examples of unsupervised online learning are online clustering, online dimensionreduction, and online anomaly detection tasks, etc.

V. Comparison of First-Order Online Learning Algorithms

Algorithm	Type of classification	Type of updates	Speed	Advantage	Limitations
Perceptron(PA)	Online binary classification	Additive updates are used	Low	Accurately classifies based on binary classifiers	1.The output can take only 0 or 1 2. It updates when classification is done wrong
Winnow(WA)	Learning from concept class	Multiplicative updates are used	Medium	Improves performance when features are irrelevant	It makes mistakes when target is calculated by performing 'OR' operation
Passive-Aggressive(PA)	Margin based learning	Classifier updates are used	Fast	1.Improves theoretical bounds 2. It updates when loss is non-zero	Updated classifier must be close to the previous classifier and should be calculated accurately
Online Gradient Descent(OGD)	Predefined learning rate scheme	Classifier updates are used	Fast	Online convex optimization	Computational cost is high

The challenging issues are:

- The most recent data point should become part of model. If that data point doesn't fit the model, the model needs to be changed or model will adapt itself based on the new data points?
- What should be the time taken to make that data point irrelevant to the model?

VI. CONCLUSION

Big data is growing and expanded in all fields like science, engineering etc. Learning from stream processing data was a challenging issue. Traditional Machine Learning techniques were not effective due to scalability, processing and continuous stream of data. In this paper, a comparative study on online learning techniques is shown for better understanding. This work has got some potential challenges for future work which will be helpful to design novel learning techniques to handle streaming big data.

REFERENCES

[1] https://kinsta.com/blog/facebook-statistics/.
[2] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 2, no. 1, pp. 1–10, 2014.
[3] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient Machine Learning for Big Data: A Review," Big Data Research, vol. 2, no. 3, pp. 87–93, Apr. 2015.
[4] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya1, Randall Wald

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2071 - 2076

and Edin Muharemagic," Deep learning applications and challenges in big data analytics" , Journal of Big Data, 2015.

[5] XW Chen, X Lin, Big data deep learning: challenges and perspectives.IEEE Access 2, 514–525 (2014).

[6] Lidong Wang, Cheryl Ann Alexander, "Machine Learning in Big Data", International Journal of Mathematical, Engineering and Management Sciences Vol. 1, No. 2, 52–61, 2016.

[7] M. Rouse, "Machine Learning Definition," 2011. http://whatis.techtarget.com/definition/machine-learning .

[8] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, and S. Li, "Incremental Learning for V-Support Vector Regression.," Neural networks : the official journal of the International Neural Network Society, vol. 67, pp. 140–50, 2015.

[9] Nataraj Dasgupta,"Hands-on techniques to implement enterprise analytics and machine learning using Hadoop, Spark, No SQL and R".

[10] PETER HARRINGTON, "Machine Learning in Action".

[11] A. L.Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, "Machine Learning With Big Data: Challenges and Approaches," in IEEE Access, vol. 5, pp. 7776-7797, 2017.

[12] R. Swathi and R. Seshadri, "Systematic survey on evolution of machine learning for big data," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp. 204-209.

[13] S. R. Sukumar, "Open research challenges with Big Data — A data-scientist's perspective," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 1272-1278.

[14] S. Mittal and O. P. Sangwan, "Big Data Analytics using Machine Learning Techniques," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 203-207.

[15] Junfei Qiu, Qihui Wu, Guoru Ding*, Yuhua Xu and Shuo Feng," A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing, 2016.

[16] Mounica Vennapusa, Srikanth Bhyrapuneni "A Comprehensive Study Of Machine Learning Mechanisms On Bigdata", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-6S2, April 2019.

[17] Junfei Qiu and Youming Sun, "A Research on Machine Learning Methods for Big Data Processing", International Conference on Information Technology and Management Innovation (ICITMI 2015).

[18] Nagwa M. Elaraby, Mohammed Elmogy, Shereif Barakat, "Deep Learning: Effective Tool for Big Data Analytics", International Journal of Computer Science Engineering (IJCSE) ISSN : 2319-7323 Vol. 5 No.05 Sep 2016 .

[19] Roheet Bhatnagar, "Machine Learning and Big Data Processing: A Technological Perspective and Review", Springer International Publishing AG, part of Springer Nature 2018.

[20] Maryam M Najafabadi,,Flavio Villanustre,Taghi M Khoshgoftaar,Randall Wald, Edin Muharemagi, "Deep learning applications and challenges in big data analytics", Najafabadi et al. Journal of Big Data (2015).

[21] Yu-Wei CD. "Machine learning with R cookbook", Packt Publishing Ltd; 2015, Mar 26.

[22] Pariwat Ongsulee, "Artificial Intelligence, Machine Learning and Deep Learning", 2017 Fifteenth International Conference on ICT and Knowledge Engineering.

[23] Wikibook, "Data Mining Algorithms In R - Wikibooks, open books for an open world", PDF generated using the open source mwlib toolkit. See http://code.pediapress.com/, 2014 14 Jul .

[24] Shai Shalev-Shwartz," Online Learning:Theory, Algorithms, and Applications",Ph.D Thesis,July 2007.[25]. Steven C. H. Hoi, Doyen Sahoo, Jing Lu, Peilin Zhao," Online Learning: A ComprehensiveStudy", Cornell Univeristy,2018.