ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Integrating Machine Learning Techniques for Big Data Analytics

Sanskruti Patel
Assistant Professor
Faculty of Computer Science and Applications, Charotar University of Science and Technology (CHARUSAT)
Changa, Gujarat, India

Abstract: The accumulation rate of data is growing tremendously than ever before. These datasets often described as Big Data as they are quite large and complex and unable to process by traditional applications. Analyzing Big Data computationally reveals many useful and interesting patterns, trends or associations. Extraction of meaningful information from massive amount of data is very much useful in many sectors. In the field of Artificial Intelligence, Machine learning is a prominent area used to uncover the hidden patterns from complex and huge datasets. This paper discussed the role of machine learning techniques in Big Data analytics and how the machine learning algorithms helps to explore massive datasets that leads towards better decision making process and prediction. It also discussed the challenges and technologies available for integrating machine learning with Big Data.

Keywords: Big Data, Machine Learning, Predictive Analysis

I. INTRODUCTION

The accumulation rate of data is growing tremendously than ever before. In digital universe, there are massive sets of data available in structured, semi-structured or unstructured form. The sources of such massive data sets are not only humans or computers but also Internet of Things (IoT) devices such as sensors, cameras, RFID's, microphones and many more. Such connected devices are generating a huge data ocean, and valuable information must be discovered from the data to help for improvement in quality of life and make our world a better place [1]. These datasets often described as Big Data as they are quite large and complex and unable to process by traditional applications. Digging and exploring such massive datasets lead towards better decision making process. Analyzing Big Data computationally reveals many useful and interesting patterns, trends or associations. However, traditional approaches are not adequate when faced with these enormous data. Organizations are interested to mine these data to gain competitive advantage and to get help in decision making.

In the field of Artificial Intelligence, Machine learning is a prominent area used to uncover the hidden patterns from complex and huge datasets. It possesses the ability of self-learning while introduced to new data. Machine learning algorithms enabled systems are highly automated and self-modifying as they continue to improve over time with minimal human intervention as they learn with more data [2][3]. Machine learning techniques are extremely powerful to make predictions on huge amounts of data [2][3]. Machine learning model first learns the knowledge from the data it is exposed to and then applies this knowledge to deliver predictions about the new data which is previously unseen [4][5][6]. Extraction of meaningful information from massive amount of data is very much useful in variety of fields including healthcare, public sector, transportation, fraud detection and many more.

This paper discussed the role of machine learning techniques in Big Data analytics and how the machine learning algorithms helps to explore massive datasets that lead towards better decision making process.

II. MACHINE LEARNING TECHNIQUES AND BIG DATA ANALYTICS

The tem Big Data was introduced by a scientist John Mashey in 1998 [8]. The term Big was arise due to the fact of massive generation of data in continuous fashion. Big Data is often characterized by five key words - Volume, Variety, Veracity, Variability and Velocity. Volume represents the extremely huge amount of data generated from various human, network and machines. Variety refers to the nature of heterogeneity in the data and the variety of data forms including structured, semi-structured and unstructured. Veracity represents the data quality that severely affects on the accuracy of the analytics. Variability refers to the data inconsistency. Velocity refers the speed of the data generation and processing. It defines the need of real time processing and availability of data [9].

On the other side, within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lends them to prediction, known as predictive analytics. These analytical models allow analysts to harvest reliable, repeatable decisions and results and uncover hidden insights through learning from historical relationships and trends in the data [7]. The following figure 1 shows an architectural diagram that describes the process of integrating machine learning techniques with Big Data sets. The process consists three major phases: Data Collection and Preparation, Model Building and Prediction.

A. Data Collection and Preparation

Due to digitalization, varieties of sources arise from where enormous data will be generated. Some of them are [10] Social

Networking sites (tweets, posts, likes, comments etc.), Internet Transactions and software applications (data input for various operations like banking, purchase, payment etc.), Mobile Devices (Call, Messages, Location etc.) and Internet of Things (IoT) Devices (sensors, internet connected hardware & software etc.).

The collected data is in raw format and required to turn in understandable format. Also, the data might be heterogeneous, incomplete, consisting impossible data combinations and out-of-range & missing values. One of the factors that affect the success of Machine learning techniques is such noisy and unreliable data. During data preprocessing, techniques like data cleaning, data reduction, data integration, data transformation may applied. It takes considerable processing time. The outcome of data preprocessing is the final training data set used during model building phase [11].

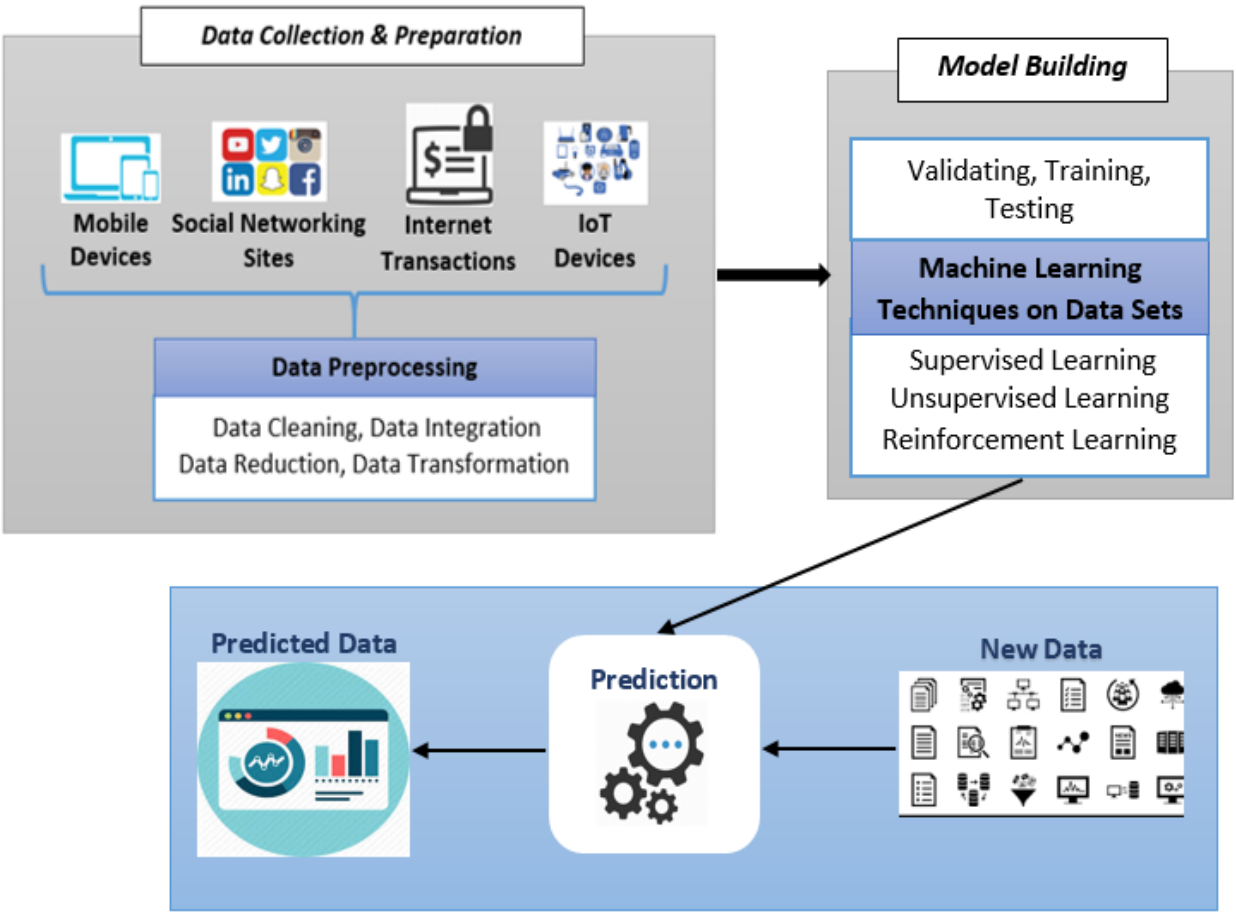

Figure 1. Integrating Machine Learning Techniques with Big Data.

B. Model Building

In this phase, a model builds and tests by applying suitable machine learning method on training data set. The input for the model building phase is the training data set. The training data set generally divided into three categories: validation, training and testing. 70% of the data is contained by training data set including 10% of the validation data set and 30% of the data is contained by test data set. Machine learning tasks typically classified into three categories: Supervised, Unsupervised and Reinforcement [12].

Supervised Learning: The data set with inputs and desired outputs is presented to an algorithm for analysis. By this means, the inputs are having pre-defined labels. The algorithm eventually learns the mapping of input to output. Supervised learning commonly uses in the applications, where historical data set is available. Supervised learning generally includes Classification and Regression. Common algorithms available for classification are Neural Network, Support Vector Machine (SVM), Decision Trees, Naïve Byes etc. Linear and Nonlinear regression are most common algorithms used for Regression [13].

Unsupervised Learning: The input data set is assigned to an algorithm without labels. The algorithm itself required to find hidden patterns or groups in the data. Unlike supervised learning, there is no availability of historical data in unsupervised learning. The most common method available for unsupervised learning is Cluster Analysis. The clusters are formed based on similarity or groups exist among data set. Common clustering algorithms [14] are Self-organizing maps, Hidden Markov models, k-Means and Hierarchical clustering etc.

Reinforcement Learning: In reinforcement learning, a software agent takes actions in a specified environment for getting maximum rewards. Software agent automatically decides an ideal behavior to maximize its performance while considering a specific context [15]. It leads to trial and error learning process. After choosing a specified algorithm, a model is first trained and then tested. The testing process continues until desired accuracy obtained.

C. Prediction

Machine learning models possess the ability of generalization and it is applicable to real world problem. Once a model is tested and verified, it can be used with new data which has been never seen before by the model. When new data comes, a model applied to get predicated data. It has been used widely in many fields including biotechnology, supply chain, medical diagnosis and behavior analysis.

Most general examples of machine learning applications on Big Data are weather forecasting, fraud detection, medical imaging, online recommendations, social media analysis, healthcare marketing, disaster management etc.

III. CHALLENGES AND ISSUES

Despite of technical advancement in the field of machine learning and Big Data, there are many significant challenges occurred which may have greater impact on success of applying machine learning techniques for Big Data analytics. Some of these are mentioned in this section.

Learning from Heterogeneous, Uncertain and Incomplete data

As data are collected from different sources, it is quite common that they are heterogeneous, uncertain and incomplete. Data can be in form of structured, semi-structured or even in unstructured form. Also, the amount of collected data is now a days enormous. Such data may have greater impact on success of machine learning algorithms as learning from these data is quite complex. Semantic and ontological representation of data possibly helps to derive the common meaning of data [23].

Data Privacy and Security

The privacy and security of data is one of the biggest challenges while applying machine learning algorithms as to process massive amount of data, distributed parallel programming framework is used. In distributed parallelism architecture, different parts of the data may handle by different owners. At that time, proper policy for trust and right management is necessary to implement.

Scaling with Big Data

Most of the machine learning algorithms required to have training dataset in the main memory. In case of Big data, it is quite complex. Distributed computing and Online Learning are not sufficient for learning from massive datasets as the size of the data is too big and too much time for training is required for sequential online learning on a single machine [21][24].

Labeling of Training Data Sets

The higher accuracy rate of most of the machine learning algorithms is depending on the training dataset provided. Training dataset is a pre-labeled dataset and in case of Big Data, labeling dataset is often expensive by considering computation time or cost. The required number of patterns depends on the extent of data and this leads to a critical issue to maintain balance between accuracy and expense [23].

IV. TECHNOLOGY PROGRESS FOR APPLICATION OF MACHINE LEARNING TECHNIQUES ON BIG DATA

There are variety of different tools and frameworks available to work on Big Data with an integration of machine learning techniques. Traditional analytical and processing tools do not work well when the volume of data is high. The essential thing in the Big Data is how to distribute the computing process and several frameworks for distributed parallel computing like Apache Hadoop, MapReduce are available for processing. However, frameworks and technologies are also needed for application of machine learning algorithms on top of these distributed computing frameworks. Following are the most common machine learning toolkits available for Big Data analytics.

Apache Mahout

Mahout is one of the well-known open source libraries available for producing scalable machine learning algorithms. Its many of the implementations use Apahce Hadoop as a processing platform. It consists classification, clustering and collaborative filtering (recommender engines) algorithms. These algorithms are implemented on top of Hadoop and uses MapReduce model. The latest stable version is 0.13.0. For classification and regression, it offers Naïve Byes, Random forest and logical regression algorithms. For Clustering, it provides algorithms like k-Means, fuzzy k-Means, spectral clustering etc. For collaborative filtering, it has algorithms for user-based filtering, item-based filtering, Matrix Factorization with ALS etc. Mahout is scalable and works well with distributed scalable processing [16][22].

MLlib

MLlib is a distributed machine learning framework that runs on top of Apache Spark – an open source cluster computing framework [22]. It is 100x faster than MapReduce and runs on Hadoop clusters and data. There are varieties of machine learning algorithms available in MLlib for classification, clustering, regression, collaborative filtering, decomposition, feature extraction, summary statistics, hypothesis testing, random data generation, etc [17]. It is possible to write applications using Java, Scala or Phython with MLlib.

RHadoop

R is an open source software environment used for statistical computing and data analysis. RHadoop is a bridging technology between R and Hadoop. RHadoop allows

distributed processing of massive data sets across clusters [18]. RHadoop come out with three R packages: rmr – offers Hadoop MapReduce functionality in R, rhdfs – provides basic connectivity to distributed file system for Hadoop and rhbase – uses for connectivity to HBase.

SAMOA

Apache SAMOA (Scalable Advanced Massive Online Analysis) is a machine learning framework works on distributed streaming processing technique [22]. It provides a collection of distributed streaming algorithms for the most of common data mining and machine learning tasks such as classification, clustering, and regression, as well as programming abstractions to develop new algorithms [19]. It is written in Java.

WEKA

WEKA is a software environment provides algorithms for data analysis and predictive modeling. WEKA 3.8 provides three different packages for distributed data mining contacting support for Hadoop and Spark called distributedWekaBase - provides base "map" and "reduce" tasks that are not tied to any specific distributed platform, distributedWekaHadoop – provides Hadoop wrappers for Weka and distributedWekaSpark – provides Spark specific wrappers for Weka [20].

V. CONCLUSION

This paper initiated with an overview of Big Data and its characteristics and then followed by an explanation of how machine learning techniques used for Big Data analytics with conceptual architecture. Moreover, challenges, issues and technological progress for integration of both the fields were also discussed to lead researchers towards research and development of more adequate techniques for integration of machine learning with Big Data analytics.

REFERENCES

[1] Wei Fan, Albert Bifet. Mining big data: current status, and forecast to the future. ACM SIGKDD Explorations Newsletter, Vol. 14, No. 2, December 2012

[2] The 10 Algorithms Machine Learning Engineers Need to Know. Top 10 Machine Learning Algorithms. https://www.dezyre.com/article/top-10-machine-learning-algorithms/202. Date Accessed: April 01, 2017

[3] Top 10 Machine Learning Algorithms. http://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html. Date Accessed: April 01, 2017

[4] T.M. Mitchell. Machine Learning. McGraw-Hill, 1997

[5] O. Chapelle, B. Cholkopf, A. Zien. SemiSupervised Learning (Adaptive Computation and Machine Learning Series). MIT Press, Cambridge, 2006.

[6] Prof. Kanchan M. Tarwani, Prof. Saleha S. Saudagar, Prof. Harshal D. Misalkar. Machine Learning in Big Data Analytics: An Overview. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 4, April 2015

[7] Joe Barron, Data Monetization. Internet of Things and Data Science as a Service – Part 2, November 01, 2016. http://www.nrconsults.com/blogs/587/. Date Accessed: April 30, 2017

[8] F. Diebold. On the Origin(s) and Development of the Term ”Big Data”. Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.

[9] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, 2001.

[10] Big Data Baby Steps. https://www.solstice.com/blog/big-data-baby-steps. Date Accessed: March 30, 2017.

[11] S. Kotsiantis, D. Kanellopoulos, P. Pintelas. Data Preprocessing for Supervised Learnin. International Journal of Computer Science, 2006, Vol. 1 No. 2, pp. 111–117.

[12] Russell, Stuart; Norvig, Peter. Artificial Intelligence: A Modern Approach (2nd ed.), Prentice Hall. ISBN 978-0137903955.

[13] Supervised Learning. https://www.mathworks.com/discovery/supervised-learning.html. Date Accessed: April 30, 2017.

[14] Unsupervised Learning. https://www.mathworks.com/discovery/unsupervised-learning.html. Date Accessed: April 30, 2017.

[15] Reinforcement Learning. http://reinforcementlearning.ai-depot.com/. Date Accessed: April 30, 2017.

[16] Apache mahouts. http://mahout.apache.org/. Date Accessed: May 02, 2017.

[17] MLlib: Scalable Machine Learning on Spark. https://stanford.edu/~rezab/sparkworkshop/slides/xiangrui.pdf. Date Accessed: May 02, 2017.

[18] Hadoop and R with RHadoop. http://www.adaltas.com/en/2012/05/19/hadoop-and-r-is-rhadoop/. Date Accessed: May 15, 2017.

[19] Morales, Gianmarco De Francisci, and Albert Bifet. "SAMOA: scalable advanced massive online analysis." Journal of Machine Learning Research. Vol. 16, pp.149-153, 2015

[20] Mining Big Data using Weka 3. http://www.cs.waikato.ac.nz/ml/weka/bigdata.html. Date Accessed: May 15, 2017.

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

[22] Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter, Tawfiq Hasanin. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, Vol. 2, No. 1, 2015

[23] J. Qiu, Q. Wu, G. Ding, Y. Xu, S. Feng. A survey of machine learning for big data processing, EURASIP J. Adv. Signal Process, pp. 1–16, 2016

[24] Junfei Qiu, Youming Sun. A Research on Machine Learning Methods for Big Data Processing. International Conference on Information Technology and Management Innovation (ICITMI 2015). www.atlantis-press.com/php/download_paper.php?id=25840020.