TRANSFORMATION OF BIG DATA THROUGH MACHINE LEARNING AND IT'S APPLICATIONS

Gone Prem Kumar
M.Tech Student
Ramchandra college of Engineering
Eluru

ABSTRACT

Both Sciences and Industry are towards a data revolution. And this has led to a complete data of new formats and unparalleled data bases. Such an increase in huge amount of data have given rise to an opportunity for Machine Learning and Bigdata to come concurrently and to develop Machine Learning methods that have the capability to hold present data types and for navigation of large amount of information with minimal or no human intervention. By implementing fast and effective algorithms and information driven models for processing of data, Machine Learning is capable to give faultless results. Today Machine Learning is being vigorously utilized in a wide range of areas than we anticipate. A pure Machine Learning process, the more data provided to the system, the more it can learn from it, returning the results that are looking for, and that's why it works well with Bigdata. Without it, the Machine Learning can't keep running at its at most level and this is because of the way that with less information, the machine has less examples to gain from, and subsequently its results may be influenced. This paper gives the survey on applications and challenges of Machine Learning techniques, advanced learning methods towards Bigdata.

KEYWORDS: *Machine Learning, Bigdata, Deep Learning, Neural Networks.*

INTRODUCTION:

It is obvious that we are living in a data deluge era, evidenced by the phenomenon that enormous amount of data have been being continually generated at unprecedented and ever increasing scales. Large-scale data sets are collected and studied in numerous domains, from engineering sciences to social networks, commerce, biomolecular research, and security. Particularly, digital data, generated from a variety of digital devices, are growing at astonishing rates. In 2011, digital information has grown nine times in volume in just 5 years and its amount in the world will reach 35 trillion gigabytes by 2020. Therefore, the term "Big Data" was coined to capture the profound meaning of this data explosion trend.

To clarify what the big data refers to, several good surveys have been presented recently and each of them views the big data from different perspectives, including challenges and opportunities, background and research status, and analytics platforms. Among these surveys, a comprehensive overview of the big data from three different angles, i.e., innovation, competition, and productivity, was presented by the McKinsey Global Institute (MGI). Besides describing the fundamental techniques and technologies of big data, a number of more recent studies have investigated big data under particular context. For example, gave a brief review of the features of big data from Internet of Things (IoT). Some authors also analyzed the new characteristics of big data in wireless networks, e.g., in terms of 5G.

Over the past decade, machine learning techniques have been widely adopted in a

number of massive and complex data-intensive fields such as medicine, astronomy, biology, and so on, for these techniques provide possible solutions to mine the information hidden in the data. Nevertheless, as the time for big data is coming, the collection of data sets is so large and complex that it is difficult to deal with using traditional learning methods since the established process of learning from conventional datasets was not designed to and will not work well with high volumes of data. For instance, most traditional machine learning algorithms are designed for data that would be completely loaded into memory, which does not hold any more in the context of big data. Therefore, although learning from these numerous data is expected to bring significant science and engineering advances along with improvements in quality of our life, it brings tremendous challenges at the same time.

The goal of this paper is twofold. One is mainly to discuss several important issues related to learning from massive amounts of data and highlight current research efforts and the challenges to big data, as well as the future trends. The other is to analyze the connections of machine learning with modern signal processing (SP) techniques for big data processing from different perspectives. The main contributions of this paper are summarized as follows:

- We first give a brief review of the traditional machine learning techniques, followed by several advanced learning methods in recent researches that are either promising or much needed for solving the big data problems.
- We then present a systematic analysis of the challenges and possible solutions for learning with big data, which are in terms of the five big data characteristics such as volume, variety, velocity, veracity, and value.
- We next discuss the great ties of machine learning with SP techniques for the big data processing.
- We finally provide several open issues and research trends.

- The remainder of the paper, as the roadmap given in Fig. 1 shows, is organized as follows. In this we start with a review of some essential and relevant concepts about machine learning, followed by some current advanced learning techniques. Section provides a comprehensive survey of challenges bringing by big data for machine learning, mainly from five aspects. The relationships between machine learning and signal processing techniques for big data processing are presented in this Section gives some open issues and research trends. Conclusions are drawn in Section.

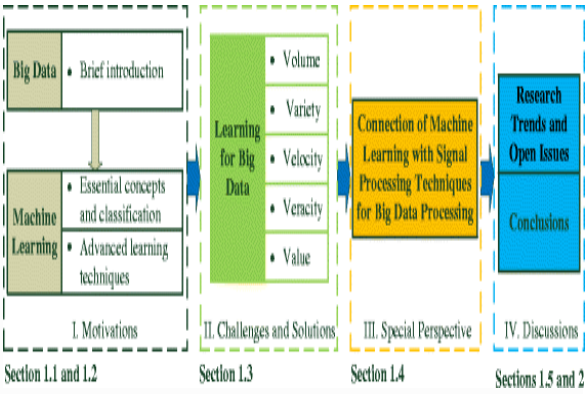

FIG 1-Machine Learning and signal processing techniques for big data

LITERATURE REVIEW:

Ming Ke, et.al.., (2014), In recent years, "Big Data" has attracted increasing attention. It has already proved its importance and value in several areas, such as aerospace research, biomedicine, and so on. In "Big Data" era, financial work which is dominated by transaction, business record, business accounting and predictions may spring to life. This paper makes an analysis about what change that "Big Data" brings to Accounting Data Processing, Comprehensive Budget Management, and Management Accounting through affecting the idea, function, mode, and method of financial management. Then the paper states the challenges that "Big Data" brings to enterprise aiming to illustrate that only through fostering strengths and circumventing weaknesses can an enterprise remain invincible in "Big Data" era.

11. Maryam M, (2015), Customer retention in telecommunication is one of the prime issue in customer relationship management (CRM). The primary focus of CRM is on existing customer as it is difficult to acquire new customers. The main goal of churn prediction is to classify customers into churner & non-churner. Towards this, deep learning as they are equipped with large increasing data sizes and uncover hidden pattern insights, detects pattern, underlying risks and alert the Telecom Industry about customer behavior with a better accuracy as compared to the traditional machine learning methods. In this paper, Deep learning by Convolutional Neural Network (CNN) is implemented for churn prediction and it showed good performance in terms of accuracy**.**

Wei Dai, et.al.., (2014), Recent years have witness the development of cloud computing and the big data era, which brings up challenges to traditional decision tree algorithms. First, as the size of dataset becomes extremely big, the process of building a decision tree can be quite time consuming. Second, because the data cannot fit in memory any more, some computation must be moved to the external storage and therefore increases the I/O cost. To this end, we propose to implement a typical decision tree algorithm, C4.5, using MapReduce programming model. Specifically, we transform the traditional algorithm into a series of Map and Reduce procedures. Besides, we design some data structures to minimize the communication cost. We also conduct extensive experiments on a massive dataset. The results indicate that our algorithm exhibits both time efficiency and scalability.

Sreenivas R. Sukumar, (2014), Machine learning has found widespread

implementations and applications in many different domains in our life. However, as the big data era is coming, some traditional machine learning techniques cannot satisfy the requirements of real-time processing for large volumes of data. In response, machine learning needs to reinvent itself for big data. In this article, we provide a review of machine learning for big data processing in recent studies. Firstly, a discussion about big data is presented, followed by the analysis of the new characteristics of machine learning in the context of big data. Then, we propose a feasible reference framework for dealing with big data based on machine learning techniques. Finally, several research challenges and open issues are addressed. Introduction Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed, aiming to understand computational mechanisms by which experience can lead to improved performance. It is a highly interdisciplinary field building upon ideas from many different kinds of domains. In the past decades, machine learning has covered almost every domain of our life which is so pervasive that you probably use it dozens of times a day without knowing it. It is primarily influencing the broader world through its implementation in a wide range of applications, which has brought great impact on the science and society.

PETER HARRINGTON, Machine learning lies at the intersection of computer science, engineering, and statistics and often appears in other disciplines. As you'll see later, it can be applied to many fields from politics to geosciences. It's a tool that can be applied to many problems. Any field that needs to interpret and act on data can benefit from machine learning techniques. Machine learning uses statistics. To most people, statistics is an esoteric subject used for companies to lie about how great their products are. (There's a great manual on how to do this called How to Lie with Statistics by Darrell Huff. Ironically, this is the best-selling statistics book of all time.) So why do the rest of us need statistics? The practice of engineering is applying science to solve a problem. In engineering we're used to solving a deterministic problem where our solution solves the problem all the time. If we're asked to write software to control a vending machine, it had better work all the time, regardless of the money entered or the buttons pressed. There are many problems where the solution isn't deterministic.

METHODOLOGY:

Machine leaning is a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms. It is a highly interdisciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimal control, and many other disciplines of science, engineering, and mathematics. Because of its implementation in a wide range of applications, machine learning has covered almost every scientific domain, which has brought great impact on the science and society. It has been used on

a variety of problems, including recommendation engines, recognition systems, informatics and data mining, and autonomous control systems.

Generally, the field of machine learning is divided into three subdomains: supervised learning, unsupervised learning, and reinforcement learning. Briefly, supervised learning requires training with labeled data which has inputs and desired outputs. In contrast with the supervised learning, unsupervised learning does not require labeled training data and the environment only provides inputs without desired targets. Reinforcement learning enables learning from feedback received through interactions with an external environment. Based on these three essential learning paradigms, a lot of theory mechanisms and application services have been proposed for dealing with data tasks. For example, Google applies machine learning algorithms to massive chunks of messy data obtained from the Internet for Google's translator, Google's street view, Android's voice recognition, and image search engine. A simple comparison of these three machine learning technologies from different perspectives is given in Table 1 to outline the machine learning technologies for data processing. The "Data Processing Tasks" column of the table gives the problems that need to be solved and the "Learning Algorithms" column describes the methods that may be used. In summary, from data processing perspective, supervised learning and unsupervised learning mainly focus on data analysis while reinforcement learning is preferred for decision-making problems. Another point is that most traditional machine-learning-based systems are designed with the assumption that all the collected data would be completely loaded into memory for centralized processing. However, as the data keeps getting bigger and bigger, the existing machine learning techniques encounter great difficulties when they are required to handle the unprecedented volume of data. Nowadays, there is a great need to develop efficient and intelligent learning methods to cope with future data processing demands.

Learning types	Data processing tasks	Distinction norm	Learning algorithms
Supervised learning	Classification/Regression /Estimation	Computational classifiers	Support vector machine
		Statistical classifiers	Naïve Bayes
			Hidden Markov model
			Bayesian networks
		Connectionist classifiers	Neural networks
Unsupervised learning	Clustering/Prediction	Parametric	K-means
			Gaussian mixture model
		Nonparametric	Dirichlet process mixture model
			X-means
Reinforcement learning	Decision-making	Model-free	Q-learning
			R-learning
		Model-based	TD learning
			Sarsa learning

Advanced learning methods

In this subsection, we introduce a few recent learning methods that may be either promising or much needed for solving the big data problems. The outstanding characteristic of these methods is to focus on the idea of learning, rather than just a single algorithm.

1. **Representation Learning:** Datasets with high-dimensional features have become increasingly common nowadays, which challenge the current learning algorithms to extract and organize the discriminative information from the data. Fortunately, representation learning, a promising solution to learn the meaningful and useful representations of the data that make it easier to extract useful information when building classifiers or other predictors, has been presented and achieved impressive performance on many dimensionality reduction tasks. Representation learning aims to achieve that a reasonably sized learned representation can capture a huge number of possible input configurations, which can greatly facilitate improvements in both computational efficiency and statistical efficiency.

   There are mainly three subtopics on representation learning: feature selection, feature extraction, and distance metric learning. In order to give impetus to the multidomain learning ability of representation learning, automatic representation learning, biased representation learning, cross-domain representation learning, and some other related techniques have been proposed in recent years. The rapid increase in the scientific activity on representation learning has been accompanied and nourished by a remarkable string of empirical successes in real-world applications, such as speech recognition, natural language processing, and intelligent vehicle systems.

1. **Deep learning:** Nowadays, there is no doubt that deep learning is one of the hottest research trends in machine learning field. In contrast to most traditional learning techniques, which are considered using shallow-structured learning architectures, deep learning mainly uses supervised and/or unsupervised strategies in deep architectures to automatically learn hierarchical representations. Deep architectures can often capture more complicated, hierarchically launched statistical patterns of inputs for achieving to be adaptive to new areas than traditional learning methods and often outperform state of the art achieved by hand-made features. Deep belief networks (DBNs) and convolutional neural networks (CNNs) are two mainstream deep learning approaches and research directions proposed over the past decade, which have been well established in the deep learning field and shown great promise for future work. Due to the state-of-the-art performance of deep learning, it has attracted much attention from the academic community in recent years such as speech recognition, computer vision, language processing, and information retrieval. As the data keeps getting bigger, deep learning is coming to play a pivotal role in providing

predictive analytics solutions for large-scale data sets, particularly with the increased processing power and the advances in graphics processors. For example, IBM's brain-like computer and Microsoft's real-time language translation in Bing voice search have used techniques like deep learning to leverage big data for competitive advantage.

2. **Distributed and parallel learning:** There is often exciting information hidden in the unprecedented volumes of data. Learning from these massive data is expected to bring significant science and engineering advances which can facilitate the development of more intelligent systems. However, a bottleneck preventing such a big blessing is the inability of learning algorithms to use all the data to learn within a reasonable time. In this context, distributed learning seems to be a promising research since allocating the learning process among several workstations is a natural way of scaling up learning algorithms. Different from the classical learning framework, in which one requires the collection of that data in a database for central processing, in the framework of distributed learning, the learning is carried out in a distributed manner. In the past years, several popular distributed machine learning algorithms have been proposed, including decision rules, stacked generalization, meta-learning, and distributed boosting. With the advantage of distributed computing for managing big volumes of data, distributed learning avoids the necessity of gathering data into a single workstation for central processing, saving time and energy. \Similar to distributed learning, another popular learning technique for scaling up traditional learning algorithms is parallel machine learning. With the power of multicore processors and cloud computing platforms, parallel and distributed computing systems have recently become widely accessible.
3. **Transfer learning:** A major assumption in many traditional machine learning algorithms is that the training and test data are drawn from the same feature space and have the same distribution. However, with the data explosion from variety of sources, great heterogeneity of the collected data destroys the hypothesis. To tackle this issue, transfer learning has been proposed to allow the domains, tasks, and distributions to be different, which can extract knowledge from one or more source tasks and apply the knowledge to a target task. The advantage of transfer learning is that it can intelligently apply knowledge learned previously to solve new problems faster.

Based on different situations between the source and target domains and tasks, transfer learning is categorized into three sub settings: inductive transfer learning, transductive transfer learning,

and unsupervised transfer learning. In terms of inductive transfer learning, the source and target tasks are different, no matter when the source and target domains are the same or not. Transductive transfer learning, in contrast, the target domain is different from the source domain, while the source and target tasks are the same. Finally, in the unsupervised transfer learning setting, the target task is different from but related to the source task. Furthermore, approaches to transfer learning in the above three different settings can be classified into four contexts based on "What to transfer," such as the instance transfer approach, the feature representation transfer approach, the parameter transfer approach, and the relational knowledge transfer approach. Recently, transfer learning techniques have been applied successfully in many real-world data processing applications, such as cross-domain text classification, constructing informative priors, and large-scale document classification.

4. **Active learning:** In many real-world applications, we have to face such a situation: data may be abundant but labels are scarce or expensive to obtain. Frequently, learning from massive amounts of unlabeled data is difficult and time-consuming. Active learning attempts to address this issue by selecting a subset of most critical instances for labeling. In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data. It can obtain satisfactory classification performance with fewer labeled samples via query strategies than those of conventional passive learning. There are three main active learning scenarios, comprising membership query synthesis, stream-based selective sampling and pool-based sampling. They have been studied extensively in the field of machine learning and applied to many data processing problems such as image classification and biological DNA identification.
5. **Kernel-based learning:** Over the last decade, kernel-based learning has established itself as a very powerful technique to increase the computational capability based on a breakthrough in the design of efficient nonlinear learning algorithms. The outstanding advantage of kernel methods is their elegant property of implicitly mapping samples from the original space into a potentially infinite-dimensional feature space, in which inner products can be calculated directly via a kernel function. For example, in kernel-based learning theory, data x in the input space XX is projected onto a potentially much higher dimensional feature space FF via a nonlinear mapping Φ as follows:

$$\Phi:X\rightarrow F,x\mapsto\Phi(x)\Phi:X\rightarrow F,x\mapsto\Phi(x) \quad (1)$$

In this context, for a given learning problem, one now works with the mapped data $\Phi(x) \in \mathcal{F}$ instead of $x \in X x \in X$. The data in the input space can be projected onto different feature spaces with different mappings. The diversity of feature spaces gives us more choices to gain better performance, while in practice, the choice itself of a proper mapping for any given real-world problem may generally be nontrivial. Fortunately, the kernel trick provides an elegant mathematical means to construct powerful nonlinear variants of most well-known statistical linear techniques, without knowing the mapping explicitly. Indeed, one only needs to replace the inner product operator of a linear technique with an appropriate kernel function k (i.e., a positive semi-definite symmetric function), which arises as a similarity measure that can be thought as an inner product between pairs of data in the feature space. Here, the original nonlinear problem can be transformed into a linear formulation in a higher dimensional space $\mathcal{F}$ with an appropriate kernel k :

$k(x, x') = \langle \Phi(x), \Phi(x') \rangle F, \forall x, x' \in X$ $k(x, x') = \langle \Phi(x), \Phi(x') \rangle F, \forall x, x' \in X$ (2)

The most widely used kernel functions include Gaussian kernels and Polynomial kernels. These kernels implicitly map the data onto high-dimensional spaces, even infinite-dimensional spaces. Kernel functions provide the nonlinear means to infuse correlation or side information in big data, which can obtain significant performance improvement over their linear counterparts at the price of generally higher computational complexity. Moreover, for a specific problem, the selection of the best kernel function is still an open issue, although ample experimental evidence in the literature supports that the popular kernel functions such as Gaussian kernels and polynomial kernels perform well in most cases

At the root of the success of kernel-based learning, the combination of high expressive power with the possibility to perform the numerous analyses has been developed in many challenging applications, e.g., online classification, convexly constrained parameter/function estimation, beamforming problems, and adaptive multiregression. One of the most popular surveys about introducing kernel-based learning algorithms is, in which an introduction of the exciting field of kernel-based learning methods and applications was given.

The critical issues of machine learning for big data:

In spite of the recent achievement in machine learning is great as mentioned in Section 1.2, with the emergence of big data, much more needs to be done to address many significant challenges posted by big data. In this section, we present a discussion about the critical issues of machine learning techniques

for big data from five different perspectives, as described in Fig. 2, including learning for large scale of data, learning for different types of data, learning for high speed of streaming data, learning for uncertain and incomplete data, and learning for extracting valuable information from massive amounts of data. Also, corresponding possible remedies to surmount the obstacles in recent researches are introduced in the discussion.

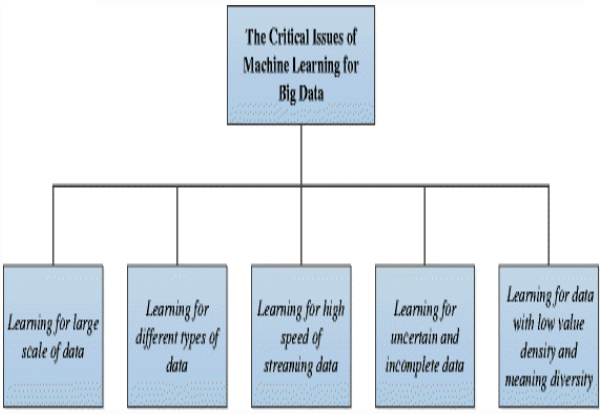

FIG 2: The critical issues of machine learning for big data

CONCLUSIONS AND DISCUSSIONS:

This paper has provided a systematic study of the traditional Machine Learning mechanisms and their comparative study when applied with Bigdata. As these algorithms are not inclusive, advanced learning methods are discussed with their issues and challenges. Various techniques have been implemented to process Machine Learning algorithms to access large scale data such as MapReduce and distributed frameworks such as Hadoop. Advanced methods include several mechanisms in which Deep learning has the potential to conquer the difficulties of Machine Learning with Big Data. Deep learning has the capability in dealing and learning problems found in huge volumes of input data in spite of having few challenges. The progressive learning and extraction of various levels of data abstractions in Deep Learning gives a certain degree of explication for Big Data Analytics. Highlight a section that you want to designate with a certain style, and then select the appropriate name on the style menu. The style will adjust your fonts and line spacing. Do not change the font sizes or line spacing to squeeze more text into a limited number of pages. Use italics for emphasis; do not underline.

REFERENCES:

1. Lidong Wang, "Data Mining, Machine Learning and Big Data Analytics" , International Transaction of Electrical and Computer Engineers System, 2017, Vol. 4, No.2, 55-61

2. Ming Ke, Yuxin Shi, "Big Data, Big Change: In the Financial Management", Open Journal of Accounting, 2014, 3, 77-82

3. Yuhua Xu and Shuo Feng ,Junfei Qiu, Qihui Wu, Guoru Ding, "A survey of machine learning for big data Processing", EURASIP Journal on Advances in Signal Processing (2016) 2016:67

4. 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)

5. Sreenivas R. Sukumar, "A Research on Machine Learning Methods for Big Data Processing", Conference Paper August 2014.

6. Omar Y. Al-Jarrah , Paul D. Yoo,Sami Muhaidat,George K. Karagiannidis,Kamal Taha, "Efficient Machine Learning for Big Data: A Review", Big Data Research2(2015) 87–93

7. 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

8. 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

9. Alexandra L'Heureux, Katarina Grolinger, Hany F. ElYamany, Miriam A. M. Capretz, "Machine Learning with Big Data: Challenges and Approaches", IEEE Access April 2017

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

11. 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)

12. Pwint Phyu Khine, Wang Zhao Shun1, "Big Data for Organizations: A Review", Journal of Computer and Communications, 2017, 5, 40-48

13. Anushree Priyadarshini and SonaliAgarwal, "A Map Reduce based Support Vector Machine for Big Data Classification", International Journal of Database Theory and Application Vol.8, No.5 (2015), pp.77-98

14. Tamer Tulgar, Ali Haydar and Ibrahim Ersan, "A Distributed K-Nearest Neighbor Classifier for Big Data", BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 6, No. 2, April 2018

15. Wei Dai, Wei Ji, "A Map Reduce approach of c4.5 Decision tree Algorithm", International Journal of Theory and Application; vol 7 no.1(2014), pp 49-60

16. Kairan Sun, Xu Wei, Gengtao Jia, Risheng Wang, and Ruizhi Li, "Large-scale Artificial Neural Network:MapReduce-based Deep Learning", arXiv:1510.02709v1 [cs.DC] 9 Oct 2015

17. .Available [online] https://www.forbes.com/sites/bernardmarr/2018

18. PETER HARRINGTON, "Machine Learning in Action", ISBN 9781617290183, Printed in the United States of America

19. Annina Simon, Mahima Singh Deo, Mahima Singh Deo, S. Venkatesan, D.R. Ramesh Babu,D.R. Ramesh Babu, "An Overview of Machine Learning and its Applications", International Journal of Electrical Sciences & Engineering (IJESE);Vol1, Issue 1; 2015 pp. 22-24