IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Review of Machine Learning Method for Resolving Issues of Big Data Analytics

M. Sharma[1*], I.S. Sohal[2], R.M.Singh[3], A. Wadhwa[4], D. Garg[5]

[1,2,3,4,5]Dept. of Computer Science, Chitkara University, Punjab, India

Corresponding Author: mudita171492.cse@chitkara.edu.in

DOI: https://doi.org/10.26438/ijcse/v7i4.559563 | Available online at: www.ijcseonline.org

Accepted: 11/Apr/2019, Published: 30/Apr/2019

***Abstract*—**In this technology bound era, data analytics is a decisive way to deal with this enormous amount of data that is getting collected from various sources such as social media, banking, healthcare etc. With the growing volume of this data, it has been getting more and more difficult to analyze the same with the existing techniques. This is where the concept of Machine Learning (ML) has turned out to be an indispensable way for giving this data an intelligent structure i.e. by sorting the clusters of data into data sets and drawing associations from previous information. However, the traditional machine learning methods are not helpful in manipulating the data in a way that we require as we are advancing in these various fields involving big data. In our research we have reviewed the various ML algorithms and learning paradigms for handling the big data problems by associating them with the challenges of the 6 big data dimensions- Volume, Veracity, Velocity, Variety, Visualization and Value. We have studied the similar approach of research given by Alexandera *et al.* and Gandomi and Haider. Adding on to their findings and methods we have considered two more V's – Visualization and Value and associated their characteristic challenges with the ML methods. We have mentioned the use of ML in preserving the privacy and security of the data as securing the data being generated is also a significant problem that needs to be addressed.

***Keywords*---**Big data analytics, Machine Learning, V's of big data, algorithms, learning paradigms

I. INTRODUCTION

With the augmentation of technology in the contemporary society, where social media and the internet have made it possible to get a plethora of information at one's fingertips, there has been a marked rise in number of people using the internet and other electrical gadgets. The data thus obtained is in huge numbers and is referred as Big Data. This data is gathered from various sources such as mobile phone applications, emails, videos, click streams [17], social media, health sectors [2], education departments [3]. Analyzation of this data to find the possible patterns and correlations is termed as big data analysis. The characteristics of big data can be divided into categories which are known as the Vs of big data. Most commonly discussed ones are volume, velocity, variety, veracity and volume [11]. These dimensions make it easy to narrow down the challenges that big data faces. To understand the characteristics of the data, some statistical and geometric patterns need to be extracted [11]. Machine Learning is a method to tackle this problem by training the data and giving it an intelligent structure. As the technology has advanced, the conventional methods of machine learning are no longer ideal. Many researches have been presented to overcome the challenges using various approaches. In 2014 Sukumar [5] in his research highlighted three important requirements for working with large datasets. They include scheming adaptive and highly scalable planning, developing ability and understanding analytical data elements before applying algorithms. In 2015, Najafabadi et al. [6] showed how deep learning could solve the general problems of machine learning for big data i.e. unstructured data, noisy data, high scalability of algorithms, streaming data, unlabeled data etc. presented their research focused on signal processing in machine learning. They focused on the five critical problems that include learning of large scale data, types of data, streaming data, uncertain data, low value data and related these to the characteristics of big data. Although these research papers are really informative, the lack of correlation of the challenges with their specific solution makes it difficult to draw conclusions. In 2105 Al-Jarrah et al. [4] surveyed machine learning for Big Data and their focus was to improve the effectiveness of wide-ranging systems and also the new algorithmic approaches for reducing memory being used. Al- Jarrah et al. mentioned the statistical aspect but the methods for minimizing computational complexity were not evaluated [9]. The approach used by Gandomi and Haider [10] to categorize the challenges according to the big data characteristics makes it easy to understand loopholes in big data analytics. However, they have not discussed the machine learning methods for the

problems. Alexandera et al. [9] have used this approach in their research, of categorizing the characteristics of the big data V's, and how different machine learning methods can be used to address the challenges. We have used this same approach in our research work Alexandera et al. have discussed about the 4 V's of big data- volume, variety, velocity, veracity, but there are two other important dimensions also that are visualization [33] and value [7]. We have studied their characteristic challenges also and associated them with the machine learning methods. Also security and privacy still remain an issue in Big data analytics. We have surveyed the possible machine learning algorithms resolving this issue which therefore also opens doors for future research work.

II. METHODOLOGY

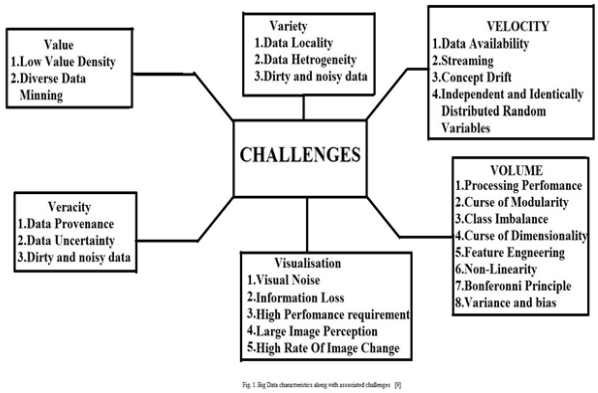

Fig. 1

A. Analytical Manipulation of Data

For data analytics using manipulation there are three methods.

These methods are shown in the fig2:

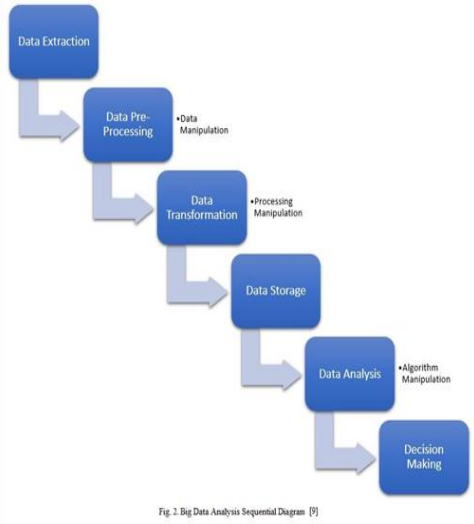

Fig. 2

1. *Data manipulations*

- The process of organizing the data in order to make it easier to read, differentiate and locate is data manipulation. In order to do that. Two major aspects affecting the machine learning conventional methods in big data are high dimensionality of datasets and huge amount of sample datasets [9]. Therefore, three types of data manipulations – dimensionality reduction, instance selection and data clearing. [9].
- *Dimensionality reduction* - To map high measurement space to lower measurement space is called dimensionality reduction [9]. Techniques used for doing this are Hessian LLE, Isomap, PCA, locally linear embedding (LLE), and Laplacian Eigenmaps. [26]. Another method is Autoencoders for encoding datasets [25]. Dimensionality reduction aims to tackle the problems of large image perception, processing challenges and the dimensionality problems.
- *Instance selection*- Instance selection reduces size of the dataset and acts as a solution to imbalance in class [9]. Its approaches include progressive sampling, using domain knowledge, random selection, cluster sampling and genetic algorithm-based selection [12]. It tackles the problems of processing performance and modularity cure.
- *Data cleaning*- It refers to the processing of wrong or inaccurate information in data. The methods used for this are smoothing filters and wavelet transforms [13]. Even Autoencoders can be used in order to remove the noise from a corrupted input data [9]. It challenges the problems of dirty and noisy data.

2. *Processing manipulation*

This method centers on how data is stored and worked on [9]. The 3 phases in which the processing happens (is shown in fig 2) are data transformation, data storage, and data estimation. In order to capture the origin of the data so that it can be modified various techniques are used for e.g. Reduce and Map Provenance (RAMP) [14] but these methods have certain computational challenges. Parallelization is a good solution for the processing manipulations. Algorithms such as brute-force search and genetic algorithms, are parallel in their approach and may show a lot of performance improvement. Based on this parallelization, researchers have developed two methods- horizontal and vertical scaling paradigms.

- *Vertical scaling*- In this, graphic processing units (GPUs) are used as a machine learning approach in big data. GPUs were firstly structured for picture display, image processing, matrix operations, and vector operations are suited for such systems [9]. If they can be parallelised other machine learning methods can be implemented on them. However, other algorithms can

be implemented on a GPU if they can be parallelized to an acceptably high level.

- *Horizontal scaling*- In this paradigm, processing is distributed over interlinked sites. There are further 2 divisions for this method - batch oriented systems and stream-oriented systems [9].

B. *Learning paradigms*

1. *Deep learning*

Deep learning is used for creating categorisations of data by virtue of being a learning paradigm that is used to produce data representations, the most ubiquitous one being that of neural nets. [31] It changes data into abstract depictions that enables the features to be learnt [9]. This method is uses to process data layer by layer. As data passes through the hidden layers, nonlinear transformations are applied [9]. Every layer tries to diverge out the variation circumstances [9]. Autoencoders, Restricted Boltzmann Machine [19] and deep belief networks [20] are used as the building blocks of deep learning [9] [19].

2. *Online learning*

It can be defined as the ability to maintain continuous updates of models of data and it requires meticulous streaming of data. it consists of two major sub methods that are facilitating the simulation of continuous processing through the means of micro blades, or utilising processes to process data continuously [31]. This paradigm uses "learn as-you-go" approach [9] which lessens the computational load and improves processing performance. This also helps with the imprecation of modularity, real-time processing challenge and data locality.

3. *Transfer learning*

It is particularly focused towards approaching the problem of insufficient data for the domain in question [31]. By this method we basically target particular domains of the dataset. This method is useful when data size in the domain is insufficient. To ensure that this paradigm is capable of being utilised for the respective domains, the metamorphosis between these different domains is elaborated on [9]. Transfer learning helps in resolving challenges related to volume (class imbalance), variety (data heterogeneity and dirty and noisy data) and veracity (data uncertainty).

Privacy and security

Three main algorithms for preserving privacy are:

- Randomisation techniques such as data perturbation [15] and matrix perturbation [16]
- Anonymization techniques like K Anonymity [1] and I diversity [18]
- Encryption techniques. For e.g. Secure Multiparty Computation and Homomorphic Encryption. [24]. This method ensures accuracy of data mining while keeping the privacy of the data as well.
- Support Vector Machines (SVM) is a very popular algorithm designed for privacy securing of data [27][28]

III. RESULTS

The tables (fig 3, fig4) show how different ML approaches aim to solve the characteristic problems of the 6 V's of big data. The '+' sign in the table indicates good remedial method and the '.' represents partial solution. In the methodology section we have surveyed and presented our finding on how different data manipulation techniques and learning paradigms aim at resolving the challenges of big data by correlating them with the measurements of big data. The 6th characteristic that is the value is considered as the output characteristic of big data. To get high value density and diversely meaningful data more research needs to be carried out to make current learning methods more flexible. The available methods- data mining and Knowledge Discovery in Database (KDD) [21] [22] [23] are used for obtaining high quality data but these methods do not focus on the diversity of the data. When it comes to od security and privacy of the data, the suggested methods (methodology sections) are very helpful. These methods allow multiple parties to collaborate globally and the large data workload can be divided into smaller jobs which is better for protection and data mining tasks [24]. Solving these characteristic challenges will further help to solve the problems being faced in the various fields which are using big data like media, entertainment, educational institutes, banks [29] , astronomy [30] etc. For example, online learning is a helpful in the stock markets as it can be used to handle stock data predictions [9]. Local learning can be used for energy consumption predictions [9].

TABLE 1: MACHINE LEARNING APPROACHES IN THE FORM OF MANIPULATIONS AND THE CHALLENGES THEY ADDRESS. [9]

APPROACHES			CHALLENGES																												
			VOLUME							VARIETY			VELOCITY			VERACITY			VISUALISATION					VALUE							
			Processing Performance	Curse of Modularity	Class Imbalance	Curse of Dimensionality	Feature Engineering	Non-linearity	Bonferroni's Principle	Variance and Bias	Data locality	Data Heterogeneity	Dirty and noisy Data	Data availability	Real-time Processing/Streaming	Concept drift	I.i.d	Data Provenance	Data Uncertainty	Dirty and Noisy Data	visual noise	Information Loss	High-quality Interface requirement	Large Image Perception	High rate of image closure	Low value density	Reverse Data Mining					
MANIPULATIONS	Data Manipulations	Dimensionality Reduction	+			+																	+	•							
		Instance Selection	+	+																											
		Data Cleaning											+							+	+										
	Processing Manipulations	Vertical Scaling	+															•													
		Hori-zontal Scaling Batch-oriented	+	+		•					+							•													
		Hori-zontal Scaling Stream-oriented	+	+										+	+			•													
	Algorithm Manipulations	Algorithm Modifications	+	•		•					+				+																
		Algorithm Mod. with new Paradigm	+	•		•					+				+																

Fig. 3

TABLE I: MACHINE LEARNING APPROACHES IN THE FORM OF LEARNING PARADIGMS ALONG WITH THE CHALLENGES THEY ADDRESS [9]

APPROACHES		CHALLENGES																												
		VOLUME								VARIETY			VELOCITY				VERACITY			VISUALISATION					VALUE				
		Processing Performance	Curse of Modularity	Class Imbalance	Curse of Dimensionality	Feature Engineering	Non-linearity	Bonferroni's Principle	Variance and Bias	Data locality	Data Heterogeneity	Dirty and noisy Data	Data availability	Real-time Processing/Streaming	Concept drift	I.i.d	Data Provenance	Data Uncertainty	Dirty and Noisy Data	visual noise	Information Loss	High performance visualization	Large Image Perception	High rate of image change	Low value density	Reverse Data Mining				
LEARNING PARADIGMS	Deep Learning					+	+				+	.						.		.			+	.						
	Online Learning	+	+	.						+		.	+	+	.	+				.		+								
	Local Learning	+	+	+					+	+												+								
	Transfer Learning			+							+	.						.	.	.										
	Lifelong Learning	+		+							+	.	+	+	.			.	.	.		+								
	Ensemble Learning	+	+												+							+								

Fig. 4

IV. CONCLUSION

To conclude, we made an attempt to survey the existing challenges in the region of big data estimation. Nowadays, machine learning is gaining much attention of the researchers for being an important tool for data scientists to overcome the data related problems. So, we reviewed the latest and improved ML procedures that may be used to solve the problems in Big data analytics. In this age of technology, where data runs in massive numbers ranging between terabytes and zettabytes [11], proper handling of this data can be very beneficial not only for the scientific society but also help in solving real world challenges. This field has a lot of scope for research as even more challenges arise from the vast amounts of data. One such problem is of concept drift. Our research also gives a direction to the future research work that can be done to the fill the gaps by presenting new methods or even by the help of these existing methods.

ACKNOWLEDGMENT

We thank Dr Anshu Singla, Nature Master Classes(NMC) and all the volunteers for their involvement in this research project. We are highly indebted to Chitkara University for their instructions and their continuous management as well as for providing their support in the completion of the project.

REFERENCES

[1] L. Sweeney, "*k-anonymity: A model for protecting privacy*," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. **10**, no. **05, 2002**

[2] A.Vinothini and Dr.S.Baghavathi Priya "*Survey of Machine Learning Methods for Big Data Applications*" International Conference on Computational Intelligence in Data Science, **2017.**

[3] Xi Fang and Juanjuan Wang, "*The Application of Big Data Technology and Method in Moral Education in Colleges and a Universities*", International conference on Intelligent Transportation, Big data and Smart City (ICITBS)

[4] S. R. Sukumar, "*Machine Learning in the Big Data Era: Are We There Yet?*" in Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining: Workshop on Data Science for Social Good (KDD 2014), **2014**.

[5] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "*Deep Learning Applications and Challenges in Big Data Analytics*," Journal of Big Data, vol. **2**, Feb. **2015**.

[6] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "*A Survey of Machine Learning for Big Data Processing*," EURASIP Journal on Advances in Signal Processing, vol. **67**, **2016**.

[7] J.Qiu et al., "A survey of machine learning for big data processing," Signal Processing for Big Data, 28 May,2016

[8] 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**, Apr. **2015**.

[9] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "*Machine Learning with Big Data: Challenges and Approaches*," IEEE Access, vol. **5**.

[10] A. Gandomi and M. Haider, "*Beyond the Hype: Big data Concepts, Methods, and Analytics*," International Journal of Information Management, vol. **35**, Apr. **2015**.

[11] S.Athmaja, M.Hanumanthapa and V.Kavitha, "*A Survey Of Machine Learning Algorithms For Big Data Analytics*," International Confrence on Innovations in Information, Embedded Communication Systems(ICIIECS),**2017**

[12] H. Liu and H. Motoda, Instance Selection and Construction for Data Mining, vol. **608**. Springer Science & Business Media, **2013**.

[13] A. Buades, B. Coll, and J. Morel, "*A Review of Image Denoising Algorithms, with a New One*," Multiscale Modeling & Simulation, vol. **4**, no. **2**, **2005**.

[14] H. Park, R. Ikeda, and J. Widom, "*RAMP: A System for Capturing and Tracing Provenance in MapReduce Workflows*.," Proceedings of the VLDB Endowment, vol. **4**, no. **12**, **2011**.

[15] R. Agrawal and R. Srikant, "*Privacy-preserving data mining*," in ACM Sigmod Record, vol. **29**, no. **2**. ACM, **2000**.

[16] K. Liu, H. Kargupta, and J. Ryan, "*Random projection-based multiplicative data perturbation for privacy preserving distributed data mining*," IEEE Transactions on knowledge and Data Engineering, vol. **18**, no. **1**, **2006**.

[17] Seref Sagiroglu and Duygu Sinanc, "*Big Data: A Review*", IEEE jornal,**2013**

[18] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*l-diversity: Privacy beyond k-anonymity*," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. **1, 2007**.

[19] R. Salakhutdinov and G. E. Hinton, "*Deep Boltzmann Machines*," in Proceedings of the International Conference on Artificial Intelligence and Statistics, **2009**.

[20] G. Hinton, "*Deep Belief Nets*," in Encyclopedia of Machine Learning, Springer, **2010,** pp. **267–269**.

[21] CW Tsai, CF Lai, MC Chiang, LT Yang, "*Data mining for the internet of things: a survey*". IEEE Commun Surv Tut **16(1), 77–97 (2014)**

[22] X Wu, X Zhu, G Wu, W Ding, "*Data mining with big data*". IEEE Trans Knowl Data Eng **26(1), 97–107 (2014)**

[23] U Fayyad, G Piatetsky-Shapiro, P Smyth, "*From data mining to knowledge discovery in databases*". AI Mag **17(3), 37–54 (1996)**

[24] Mohammed Z. Omer and Hui Gao, "*Privacy Preserving in Distributed SVM Data Mining on Vertical Partitioned Data*", 3rd International Conference on Soft Computing & Machine Intelligence, **2016**.

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

[26] C. L. Philip Chen and C. Y. Zhang, "*Data-Intensive Applications, Challenges, Techniques and Technologies: a Survey on Big Data*," Information Sciences, vol. **275**, **2014**.

[27] S. G. Teo, S. Han, and V. C. Lee, "*Privacy preserving support vector machine using non-linear kernels on hadoop mahout*," in Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on. IEEE, **2013**, pp. **941–948**.

[28] Y. Rahulamathavan, S.Veluru, R. C.-W. Phan, J. A. Chambers, and M. Rajarajan, "*Privacy-preserving clinical decision support system using gaussian kernel-based classification*," IEEE journal of biomedical and health informatics, vol. **18,** no. **1**, pp. **56–66, 2014**.

[29] Long Xu and Yihua Yan, "*Machine Learning for Astronomical Big Data Processing*", IEEE,**2017**

[30] A.Vinothini, S.Baghavati, "*Survey of Machine Learning for Big Data Applications*," International Confrence on Computational Intelligence in Data Science(ICCIDS),**2017**

[31] C.Augenstein, N.Spangenberg and Bogdan Franczyk, "*Applying Machine Learning to Big Data Streams*,"4th IEEE International Conference on Soft Computing and Machine Intelligence,**2017**

