IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Review on Machine Learning Techniques for Big Data Management and Open Research Challenges

Gagandeep Kaur[1*], Jasvir Singh[2],

[1*]MCA, Ramgarhia Institute of Engineering and Technology, Phagwara, India.
[2] MCA, Ramgarhia Institute of Engineering and Technology, Phagwara, India

[*]*Corresponding Author: gagankaur.0813@gmail.com, Tel.: +91-99150-93515*

Available online at: www.ijcseonline.org

Accepted: 10/Jul/2018, Published: 31/Jul/2018

***Abstract*—** In the recent years, the database on the cloud network is increased exponentially and this enormous volume of data is known as Big Data. The Big Data is described in five V's known as volume, velocity, variety, Veracity, and Value. Hence, efficient algorithms and architectures are required to process and store the data. In this paper, a review study on machine learning techniques for database management is done. From the study, it is found that machine-learning algorithms provide efficient data processing and storage. In the last research issues are defined which helps the other author to contribute their work in this area.

Keywords— Big Data, Machine Learning, Internet of Things, Database Management.

I. INTRODUCTION

In the last few decades due to advancement in the technology and digital data, communication on the internet is increased. To manage the digital data online cloud computing comes in picture. The high volume of data is known as Big Data. Further, the Big Data is described in various V's connected with them [1]. These V's are illuminated below.

- Volume
  This term denotes the giant amount of data is produced every second on the Internet and effective distributed systems are required to manage these amounts of data.
- Velocity
  The huge amount of data is processed randomly accessed from the user. Therefore, at which rate data is processed and get response defined velocity term.
- Variety
  The data is collected from the various sources and available in different forms.
- Veracity
  This term defined the biases or abnormalities available in the data.
- Value
  How much valuable information is retrieved from the data is known as value.

1.1 Application Areas

The following application area where the huge volume of data available and efficient algorithms are required for the process the data.

- E-Healthcare
  In the E-Healthcare network, the Big Data is produced due to keeping records of the patients, compliance and regulatory requirements [2]. Further, this information is processed and analyzed by various medical students for their research.
- Network Intrusion Detection
  The network intrusion detection systems are processed time-sensitive information and hence a large amount of data is produced every second [3]. Therefore, Big Data efficient techniques are required for real-time processing.
- Social Networks
  In the current scenario, the number of social domains are available (such as Facebook, Twitter, and Instagram etc) which produced a large amount of data on the internet network [4]. In addition, heterogeneous data is produced from this website in the form of videos, images, text files.
- E-Commerce
  Due to advancement in the internet of network E-commerce online business has been increased and the database is also increased exponentially. In the 2012, Walmart's database contains 2.5 petabytes of customer information [5].

This application shows that Big Data processing required efficient technique for better prediction and accuracy purposes.

1.2 Machine Learning in Big Data

The Big Data is complex in nature and large in volume. Hence, advanced and efficient algorithms are required to process the data. The machine learning algorithms gain attention in last few years to due to efficient processing and providing better accuracy [6].

In the Fig. 1 the machine learning for Big Data is given. The Big Data, users, domains, and systems are peripheral components for machine learning and all components are work in both directions. The Big data are the input to pre-processing in machine learning. The users interact with

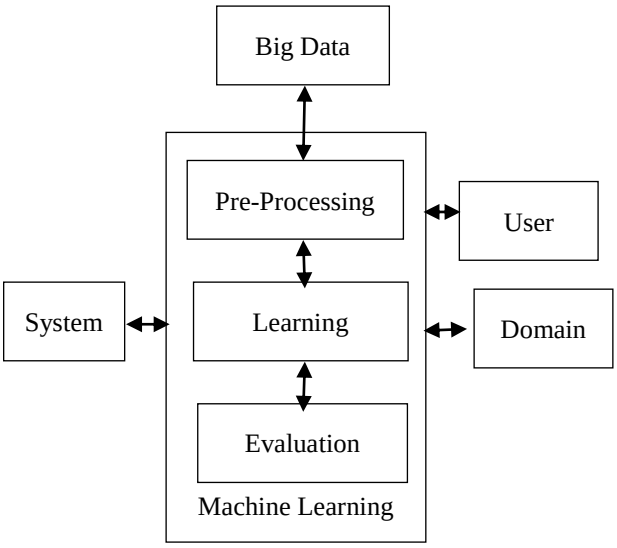

Fig. 1 Machine Learning on Big Data [7]

learning phase by providing domain information. Next, the system architecture defined how efficiently the algorithms are processed [7]. Further, the machine learning contains three steps:

- Pre-Processing
  The data is taken from Big Data unit is available in the form of unstructured, inconsistent, and noisy or incomplete. Hence, pre-processing is done which structure or fused the data and arrange in systemic way so learning phase process the data.
- Learning
  The learning process chooses learning algorithms how to process the data to get desired output.
- Evaluation
  The evaluation parameter measured the performance of the learning algorithms in terms of error estimation, statistical test.

The main contribution of this review paper to provide an efficient information of various machine-learning algorithms are deployed for the Big Data management. In addition, provided the pros, cons, research challenges of machine learning algorithms.

The rest of the paper is organized as follow, section I contains a overview of Big Data and machine learning employed on Big Data, Section II defined the literature survey is done on machine learning algorithms, section III explained the comparative analysis of machine learning algorithms, Section IV defined the research challenges, section V conclusion of the work is done.

II. LITERATURE SURVEY

In this section, a review of machine learning tools and algorithms are done.

R. Swathi and Dr. R. Seshadri [8], review the Big Data and Machine Learning. In addition, defined the different engines, tools, model and language supported in machine learning and application domains.

The machine learning algorithms are enabled the precise prediction in Big Data. The author Zhou, *et al.* [9] defined the machine learning opportunities and challenges in the Big Data. In their work, defined the framework of machine learning technique that includes various phrases such as preprocessing, learning and evolution.

Varatharajan, et al. [10], used the machine learning Linear Discriminant Analysis (LDA) with enhanced kernel based SVM algorithm for significantly identifies the Q, R, and S wave in the ECG signal. The performance analysis is done based on sensitivity, specificity, and MSE (Mean Square Error) and compared with existing LDA system.

Maillo, et al. [11], used K-Nearest Neighbour (KNN) with SPARK. The KNN is not feasible in where time and memory big concern. Hence, the author used SPARK memory to defined the training dataset. Further, to speed up the processing MapReduce scheme is used. Their implementation accelerates the data normalization, processing, computation without incurring in Hadoop startup costs.

Chen, et al. [12], proposed convolutional neural network based multi-modal disease risk prediction algorithm for hospital, structure or unstructured data available. The accuracy parameter is important in healthcare, hence the author used latent factor model for reconstructing the missing data. Their experimental results show that their technique has 94.8% accuracy with convergence speed.

[8] R. Swathi and Dr. R. Seshadri 2017. Systematic survey on evolution of Machine Learning for Big Data. International Conference on Intelligent Computing and Control Systems, pp. 204-209.

Donald C. Wunsch [13], defined that in the current scenario the complexity of the systems is increased due to exponential data is available. Therefore, fast computation based machine learning/ artificial intelligence is required. In

this paper, various research challenges for learning phase is explained.

Le, et al. [14] defined that traffic forecasting plays an important role in improving network quality and energy saving and real time outcomes required. The traffic forecasting is done based on processing historical data and key performance indicator. Hence, the author deployed machine-learning technique for traffic forecasting for different cells which includes GSM, 3G and 4G. To evaluate the performance of the proposed technique, the real dataset that collected more than 6000 cells of the real network in 2016-17 years.

Li, et al. [15], defined that in the current scenario a large amount of heterogeneous data available on the Internet of Things. This heterogeneous data is not suitable for the Convolutional Neural Network which is most preferred machine learning technique in IoT. Hence, the author proposed deep convolutional computational model and used tensor representation model to learn hierarchical features. Next, backpropagation algorithm is used for train the parameters. The technique is carried out on three dataset CUAVE, SNAE2, and STL-10 and result shows better accuracy for the proposed technique.

Shen, et al. [16], convolutional neural network (CNN) gains attention in image classification and recognition because of its performance when large amount of training images are available. But, its performance influence when the dataset is not suitable for sufficient training. Hence, extra resources are used for making it suitable for training set. In this paper, compressing sensing (CS) CNN model is proposed which consume less resource. The proposed technique is evaluated on public dataset for deep learning task using different metrics and found that CS-CNN is able to speed the training process.

Lakshmanaprabu, et al. [17], introduced a hierarchical framework for feature extraction in Social IoT big data using map-reduced framework along with a supervised classifier model. In the initial phase, the Gabor filter is used to reduce unwanted data and noise. Further, to improve the efficiency Hadoop Map Reduce has been deployed for mapping and reducing big databases. Next, Elephant Herd Optimization is used for feature selection on filter dataset. The architecture is implemented using linear kernel support vector machine to classify and predicting the efficiency. The evaluation results show that they achieved 98.2% accuracy, 85.88% specificity, and 80% sensitivity.

III. Comparative Analysis of Machine Learning Algorithms

A comparative analysis of machine learning tools and machine learning algorithms are done in the Table 1-2.

Table 1 Comparative Analysis of Machine Learning Tools [8]

Processing Engines	Associated Tools	Execution Model	Language Supported
Spark	Graphlib, Giraph, MLlib	Batch and stream	R, Python, Java
Map Reduce	Mahout	Batch	Java
Flink	SAMOA, Flink-ML	Batch and Stream	Java, Scala
Storm	SAMOA	Stream	Any

Table 2 Comparative Analysis of Machine Learning Algorithms

Machine Learning Algorithm	Pros	Cons
Support Vector Machine	1.Best when no prediction how much data available. 2.Suitable for structured and semi-structured data	1.Choosing kernel is not easy. 2. For large dataset long training time required.
Convolutional Neural Network	1.Minimize computation as compared to neural network. 2. Once trained then the prediction is fast.	The pre-processing time consuming for the train the data.
K-Nearest Neighbour	Robust to noisy and suitable for large data.	1.Need to determine the value of parameter k. 2.Computation cost is high

IV. Research Challenges

From the study and review found that in the machine learning pre-processing, learning and evaluation are the important parameter for improving the performance of the system. In this section, we highlighted some research direction on which further work can be done.

- In the E-healthcare and the traffic, forecasting in short interval of data is collected and real-time processing is required. The short intervals produced a large amount of redundant data. Hence, efficient pre-processing of data is required therefore redundant data is not stored in the database.
- The convolutional neural network (CNN) is suitable for the homogeneous dataset and for heterogeneous data pre-processing of data is done which required extra resources and memory [15]. Hence, selection of alternative machine

learning technique which suitable for homogeneous and heterogeneous data.

- In the Big Data large volume of data available which is processed and manipulate by machine learning technique to provide better accuracy with some latency. However, the number of application where real-time response is required. So, design optimized machine learning algorithms for better performance.

V. CONCLUSION

In this paper, the Big Data is described in 5 V's (volume, velocity, variety, veracity, and value). To process large amount of volume and provide velocity in data processing machine learning models are for Big Data. Hence, different machine learning techniques, tools are reviewed and comparative analysis is done in this paper. Also, defined some research direction how to improve the accuracy and give direction to other authors to contribute their work in this field.

References

[1] Athmaja, S., Hanumanthappa, M. and Kavitha, V, "A survey of machine learning algorithms for big data analytics," In Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017 International Conference, pp. 1-4, 2017.

[2] Raghupathi, W. and Raghupathi, V., "Big data analytics in healthcare: promise and potential," Health information science and systems, vol. 2, issue 1, p.3, 2014

[3] Suthaharan, S., "Big data classification: Problems and challenges in network intrusion prediction with machine learning," ACM SIGMETRICS Performance Evaluation Review, vol. 41, pp.70-73, 2014

[4] Wei Tan, Brian Blake, Iman Saleh, and Schahram Dustar, "Social-Networks-Sourced Big Data Analytics," IEEE Internet Computing, vol. 17, issue 5, pp.62-69, 2013.

[5] Edosio, U.Z., "Big data Analytics and its Application in E-commerce," Proceedings E-Commerce Technologies. University of Bradford, 2014.

[6] Singh, S.P. and Jaiswal, U.C., "Machine Learning for Big Data: A New Perspective," International Journal of Applied Engineering Research, vol. 13, issue 5, pp.2753-2762, 2018.

[7] Zhou, L., Pan, S., Wang, J. and Vasilakos, A.V., "Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp.350-361, 2017.

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

[9] Zhou, L., Pan, S., Wang, J. and Vasilakos, A.V., Machine learning on big data: Opportunities and challenges. Neurocomputing, vol. 237, pp.350-361, 2017.

[10] Varatharajan, R., Manogaran, G. and Priyan, M.K., "A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing," Multimedia Tools and Applications, vol. 77, issue 8 , pp.10195-10215, 2018.

[11] Maillo, J., Ramírez, S., Triguero, I. and Herrera, F.,. "KNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data," Knowledge-Based Systems, vol. 117, pp.3-15, 2017.

[12] Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L.,. "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, pp.8869-8879, 2017

[13] Sun, F., Huang, G.B., Wu, Q.J., Song, S. and Wunsch II, D.C.,. "Efficient and rapid machine learning algorithms for big data and dynamic varying systems," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 47, issue 10, pp.2625-2626, 2017.

[14] Le, L.V., Sinh, D., Tung, L.P. and Lin, B.S.P., "A practical model for traffic forecasting based on big data, machine-learning, and network KPIs," In Consumer Communications & Networking Conference (CCNC), 2018 15th IEEE Annual, pp. 1-4, 2018.

[15] Li, P., Chen, Z., Yang, L.T., Zhang, Q. and Deen, M.J., "Deep convolutional computation model for feature learning on big data in Internet of Things" IEEE Transactions on Industrial Informatics, vol. 14, issue 2, pp.790-798, 2018.

[16] Shen, Y., Han, T., Yang, Q., Yang, X., Wang, Y., Li, F. and Wen, H., "CS-CNN: Enabling Robust and Efficient Convolutional Neural Networks Inference for Internet-of-Things Applications," IEEE Access, vol. 6, pp.13439-13448, 2018.

[17] Lakshmanaprabu, S.K., Shankar, K., Khanna, A., Gupta, D., Rodrigues, J.J., Pinheiro, P.R. and De Albuquerque, V.H.C., "Effective Features to Classify Big Data Using Social Internet of Things," IEEE Access, vol. 6, pp.24196-24204, 2018.

Authors Profile

Mrs. Gagandeep Kaur is a MCA (Honours) from Apeejay Institute of Management Technical Campus Jalandhar. She is pursuing Ph.D from GNA University, Phagwara. She has more than 9 years of academic experience and 1 year of industrial experience and is presently working with Ramgarhia Institute of Engineering and Technology Phagwara as an Assistant Professor in the Department of Management and Computer Applications. Her areas of expertise are Data Communication and Networks, Embedded Systems, Microprocessors, Assembly Language, Database Administration, Data Warehousing, Software Engineering, System Programming, Neural Networks, Soft Computing, Artificial Intelligence, System Analysis and Design, Linux OS, Object Oriented Analysis and Design, Cloud Computing, IoT and Information Security. She has presented research papers in various national and international seminars and has been an observer, examiner and evaluator for university examinations on several occasions.

Mr. Jasvir Singh pursued Bachelor of Computer Applications from PCTE, Baddowal, Ludhiana under Punjab Technical University in 2003 and Master of Computer Applications from Main Campus, Punjabi University, Patiala. He has 5 years of academic experience and 5 years of industrial experience and is currently working as Assistant Professor in Ramgarhia Institute of Engineering and Technology, Phagwara in the department of Management and Computer Applications. His area of expertise is Data Communication and Networks, Network Programming, C, C++, Java, Python, ASP. Net, Data Base Management, E-Commerce, Software Quality and management, Linux/Unix/Mac/Windows OS. He has presented research paper in various National and International seminar and also published papers in National and International Journals and has been an observer, examiner and evaluator for university examinations on several occasions.

