IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 1, January 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 1, January 2023 ||

| DOI:10.15680/IJIRSET.2023.1201110 |

Machine Learning Algorithms for Real-Time Big Data Processing

Channappa A

Senior Scale Lecturer, Government Polytechnic, Kudligi, Karnataka, India

ABSTRACT: 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. 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. The capability to process these gigantic amounts of data in real-time with Big Data Analytics (BDA) tools and a Machine Learning (ML) algorithm carries many paybacks. However, the high number of free BDA tools, platforms, and data mining tools makes it challenging to select the appropriate one for the right task. Artificial Intelligence (AI) techniques such as machine learning and evolutionary algorithms able to provide more precise, faster, and scalable outcomes in big data analytics. Furthermore, this survey denotes the strengths and weaknesses of the selected AI-driven big data analytics techniques and discusses the related parameters, comparing them in terms of scalability, efficiency, precision, and privacy. Furthermore, a number of important areas are provided to enhance the big data analytics mechanisms in the future.

KEYWORDS: machine learning techniques, big data, Machine Learning (ML) algorithm, Big Data Analytics (BDA), efficiency, Artificial Intelligence (AI),

I. INTRODUCTION

Big data analytics is the process of examining and analyzing massive and varied data that can help organizations make more-informed business decisions, especially for uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful information. Big data analytics is one high focus of data science and there is no doubt that big data is now quickly growing in all science and engineering fields. It would be intriguing to expand the research to other organs such as the liver, lungs, and other multi-modal medical imaging modalities in the future. A new deep CNN model can be utilized for different handwriting styles. Future work can address the limitations such as identifying new features, developing classifiers with other machine learning techniques. Moreover, other works might be performed in the future to improve the classification toxic comments model's suitability for dealing with specific social media data, studying more colloquial textual data on social sites such as Twitter and Instagram and proposing deep learning models that can be enhanced with semantically rich representations to successfully extract people's ideas and attitudes. Big data has become essential as numerous organizations deal with massive amounts of specific information, which can contain useful information about problems such as national intelligence, cyber-security, biology, fraud detection, marketing, astronomy, and medical informatics. Several promising machine learning techniques can be used for big data analytics including representation learning, deep learning, distributed and parallel learning, transfer learning, active learning, and kernel-based learning. That presents one of the explanations for why this Special Issue has been named "Machine Learning Technologies for Big Data Analytics". Nevertheless, there is another explanation: practical applications need researchers, scientists, and engineers to find solutions for the big data problems consistent with current technologies and react to the demands from the near future.

II. LITERATURE REVIEW

Jincheng Zhou (2022) Big data, cloud computing, and artificial intelligence technologies supported by heterogeneous systems are constantly changing our life and cognition of the world. At the same time, its energy consumption affects the operation cost and system reliability, and this attracts the attention of architecture designers and researchers. By using the greedy method, the high-value tasks are assigned first, and then the low-value tasks are allocated. In order to verify the effectiveness and rationality of HGES, different tasks with different inputs and different comparison methods are designed and tested. Firstly, all tasks are assigned to each GPU in the system, and then the tasks are divided into high-value tasks and low-value tasks by the calculated average time value and variance value of all tasks.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 1, January 2023 ||

| DOI:10.15680/IJIRSET.2023.1201110 |

Amir H. Gandomi (2021) Data are presently being produced at an increased speed in different formats, which complicates the design, processing, and evaluation of the data. The MapReduce algorithm is a distributed file system that is used for big data parallel processing. Then, the best available machines were selected to find the optimal solution. In this manner, the optimization of resources is said to have taken place. Then, an energy efficiency-aware greedy scheduling algorithm was presented to select a position for each task to minimize the total energy consumption of the MapReduce job for big data applications in heterogeneous environments without a significant performance loss. To evaluate the performance, the LWR-EGS method was compared with two related approaches via MapReduce. Moreover, the method also reduced the execution time when compared to state-of-the-art methods.

Md Ashiqur Rahman (2020) The purpose of this research is to build a prototype system using different Machine Learning Algorithms (models) and compare their performance to identify a suitable model. This paper explores three most commonly used Machine Learning Algorithms named as Logistic Regression, Support Vector Machine and Artificial Neural Network. To conduct this research, a clinical dataset has been used. To evaluate the performance, different evaluation methods have been used such as Confusion Matrix, Stratified K-fold Cross Validation, Accuracy, AUC and ROC. The dataset contains class imbalance, so the SMOTE Algorithm has been used to balance the dataset and the performance analysis has been carried out on both sets of data. The results show that accuracy scores of all the models have been increased while training the balanced dataset.

III. BIG DATA

We now live in an era of data deluge where large volumes of data are accumulating in all aspects of our lives. Data streams coming from diverse domains contribute to the emerging paradigm of big data. It may be a great opportunity for the big data scientist amongst the vast amount and array of data. By discovering associations, analyzing patterns and predicting trends within the data, big data has the potential to change our society and improve the quality of our life. Big data typically refers to the following three types based on data sources from physical, cyber, and social worlds:

- **Nature data:** we can imagine that data coming from the nature in our earth will be a great potential data source, such as satellite data from outer space.
- **Life data:** It is a big project on the study of biological body and especially the exploration on the human body still have a lot of challenges, such as biological data.
- **Sociality data:** with the fast development of digital mobile products and network, large volumes of sociality data are generating every day in our life, such as voice abd video data.

New Features of Machine Learning with Big Data
In order to deal with the potential chanlleges posed by big data, machine learning has to possess some new properties compared with the traditional learning systems and techniques. In this section, we will highlight three aspects of abilities that are useful to deal with big data problems for machine learning techniques in detail, i.e., sparse representation and feature selection, mining structured relations, high scalability and high speed.

The types of machine learning algorithms
There are four types of machine learning algorithms: supervised, semi-supervised, un-supervised and reinforcement.

Supervised learning
In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions. The algorithm makes predictions and is corrected by the operator – and this process continues until the algorithm achieves a high level of accuracy/performance.

Under the umbrella of supervised learning fall: Classification, Regression and Forecasting.

Classification: In classification tasks, the machine learning program must draw a conclusion from observed values and determine to what category new observations belong. For example, when filtering emails as ‘spam’ or ‘not spam’, the program must look at existing observational data and filter the emails accordingly.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 1, January 2023 ||

| DOI:10.15680/IJIRSET.2023.1201110 |

Regression: In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.

Forecasting: Fore-casting is the process of making predictions about the future based on the past and present data and is commonly used to analyse trends.

Semi-supervised learning
Semi-supervised learning is similar to supervised learning, but instead uses both labelled and unlabelled data. Labelled data is essentially information that has meaningful tags so that the algorithm can understand the data, whilst unlabelled data lacks that information. By using this combination, machine learning algorithms can learn to label unlabelled data.

Unsupervised learning
Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analysing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly. The algorithm tries to organise that data in some way to describe its structure. This might mean grouping the data into clusters or arranging it in a way that looks more organised.

As it assesses more data, its ability to make decisions on that data gradually improves and becomes more refined.

Under the umbrella of unsupervised learning, fall:
Clustering: Clustering involves grouping sets of similar data (based on defined criteria). It's useful for segmenting data into several groups and performing analysis on each data set to find patterns.

Dimension reduction: Dimension reduction reduces the number of variables being considered to find the exact information required.

Reinforcement learning
Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches the machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result.

IV. RESEARCH METHODOLOGY

The methodology consists of parallel and combined processing approaches with machine learning algorithms, Bio-inspired algorithms and Fuzzy based ADNN and NBC-MVB classifier. The parallel approach is achieved through the cloud computing environment to handle the large-scale student's data. An MLR is typically shorter than a full-fledged literature review. Analytics also provides researchers with opportunities to carry out real-time analysis of activities. Data predictive analytics for instance focuses on applying tools and techniques to analyze large sets of data. The steps of innovative methodologies adopted for experimentation with large-scale student's data to yield better results in predicting the performance are discussed. We adopted MLR because its concise format makes it simple to grasp those themes in the literature, allowing more practitioners to profit from them. Initially, all the studies relevant to the current study (big data and BDA) were carefully chosen in the primary screening phase. That is because, unlike a literature review, which focuses on synthesising findings from several studies to develop conclusions on a broad area of research, an MLR focuses on a single subject or topic. It is vital to remember that the literature review and the MLR format are the same. Likewise, there is no substantial difference between the steps involved in MLR and the literature review. However, the only difference between the two is that one is broader while the other is narrower.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 1, January 2023 ||

| DOI:10.15680/IJIRSET.2023.1201110 |

V. RESULTS

The results obtained from the ADNN and NBC-MVB classifier are discussed in this section. The results of the Neural networks with the adaptive structure and Naïve Bayes classifiers with multi-variate Bernoulli model are better than the simple Neural Networks or Naïve Bayes Classifiers. The CR and FPR is compared with the rule based techniques and a considerable improvement of performance is obtained.

Table 1: CR comparison of ADNN and NBC-MVB for real time data

Sl.No	Name of theattack	CR (ADNN) (%)	CR(NBC-MVB) (%)
1	Normal	94.32	89.24
2	ICMP flood	92.00	88.62
3	Smurf flood	91.98	90.95
4	Land flood	90.98	91.96
5	TCP flood	90.99	90.68
6	UDP flood	90.00	90.49
7	Fraggle flood	90.00	88.98
8	HTTP flood	75.00	77.98
9	Session flood	78.99	79.24
	Total	86.74	85.96

Table 1 shows the comparison of CR between ADNN and NBC-MVB. Normal, ICMP flood, Smurf flood are detectable by the ADNN classifier. Unknown attack type land is highly classified by the NBC-MVB and Fraggle classified with good CR by the ADNN classifier. Application- layer DDoS attacks are classified with high CR by the NBC-MVB classifier.

Table 2: FPR comparison of ADNN and NBC-MVB for real time data

Sl.No	Name of theattack	FPR (ADNN) (%)	FPR (NBC-MVB) (%)
1	Normal	7.80	7.17
2	ICMP flood	9.61	9.45
3	Smurf flood	9.44	8.86
4	Land flood	8.82	8.49

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 1, January 2023 ||

| DOI:10.15680/IJIRSET.2023.1201110 |

5	TCP flood	9.04	8.25
6	UDP flood	7.67	7.09
7	Fraggle flood	8.47	8.07
8	HTTP flood	11.05	10.80
9	Session flood	10.59	10.42
	Total	9.92	9.37

Table 3: CR and FPR of ADNN for kddcup 99 data

Sl.No	Name of the attack	CR (%)	FPR (%)
1	Normal	94.34	6.89
2	Probe	93.91	7.64
3	Dos	90.43	7.59
4	U2R	49.52	28.76
5	R2L	38.45	24.94
	Total	86.11	17.95

Table 2 shows the comparison of FPR between ADNN and NBC-MVB. NBC-MVB FPR of Normal, ICMP flood, Smurf flood, Land flood are low when compared to the ADNN classifier. Unknown attack type Fraggle has low FPR for NBC-MVB classifier. Application-layer DDoS attacks have low FPR for the NBC-MVB classifier. Table 3 shows the CR and FPR for kddcup 99instances for ADNN classifier. U2R and R2L attacks have low CR as compared to other types of attacks.

Table 4: CR and FPR of ADNN for STCS data

Sl. No	Name of thedataset	CR (%)	FPR (%)
1	STCS data	84.23	8.93

Table 4 shows CR and FPR for STCS data using ADNN classifier. An U2R and R2L attack has low CR as compared to other types of attacks. An U2R and R2L attack has the high FPR as compared to other types of attacks. DoS attack has low FPR as compared to other types of attacks.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 1, January 2023 ||

| DOI:10.15680/IJIRSET.2023.1201110 |

VI. CONCLUSION

Machine Learning Algorithms are used to analyze and predict data efficiently. Initially, Baseline Machine Learning Algorithms such as Decision Tree, Naïve Bayes, Neural Network, and Support Vector Machines are used to analyze the student performance in standalone and virtual machines.To improve the Classification Rates (CR) Adaptive Neural Networks (ADNN) and Naïve Bayes classifiers (NBC-MVB) are used. The ADNN classifier is using the Adaptive nature of NN and the NBC-MVB is used to calculate the posterior probabilities of the data which is used to calculate the class of the attribute. Hybrid models such as NBTree and Neural Support Vector Machine (NSVM) are investigated along with it. The performance of the algorithms is analyzed based on measures such as Predictive accuracy, Precision, Recall, and F-Measure, Speed up ratio, CPU utilization, and memory utilization are evaluated. Advancement in machine learning for big data, novel machine learning methods, recently proposed learning techniques such as representation-learning, transfer-learning, deep-learning, active-learning etc. are discussed. It fails to study big data analysis techniques that are available in different sources. Without highly available systems, you will always be catching up with your data analysis and won't be able to fully capitalize on your machine learning intelligence. Furthermore, this survey introduces some interesting lines for future research.

REFERENCES

1. Shuo Feng (2016),"A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing,ISSNno:1687-6180,Vol. 2016,https://doi.org/10.1186/s13634-016-0355-x
2. Cem Tekin (2013),"Learning optimal classifier chains for real-time big data mining",51st Annual Allerton Conference on Communication,ISSNno:2836-4503,DOI:10.1109/Allerton.2013.6736568
3. Laith Abualigah (2022),"Machine Learning Technologies for Big Data Analytics",Electronics,ISSNno:1450-5843,Vol.11(3),Pages.421.https://doi.org /10.3390/electronics11030421
4. Amir H. Gandomi (2021),"Linear Weighted Regression and Energy-Aware Greedy Scheduling for Heterogeneous Big Data",Electronics,ISSNno:1450-5843,Vol.10(5),Pages.554.https://doi.org/10.3390/electronics10050554
5. Houman Homayoun (2018),"Energy-efficient acceleration of MapReduce applications using FPGAs",Journal of Parallel and Distributed Computing, ISSNno:1096-0848,Vol.119,Pages.1-17.https://doi.org/10.1016/j.jpdc.2018.02.004
6. Jincheng Zhou (2022),"Low Power Scheduling Approach for Heterogeneous System Based on Heuristic and Greedy Method",Comput Intell Neurosci., ISSNno:1687-5273,Vol.2022,doi:10.1155/2022/9598933
7. Md Ashiqur Rahman (2020),"Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease",Journal of Data Analysis and Information Processing,ISSNno:2327-7203,Vol.8,No.2,Pages. 41-68.
8. Mohammad Hasan Imam (2018),"Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System",World Journal of Engineering and Technology,ISSNno:2331-4249, Vol.6,No.4,Pages.854-873.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details