Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 9437 - 9452

Research Article

Performance Analysis of Big Data Based Mining and Machine Learning Algorithms: A Review

Ganesh Yenurkar[a], Dr. Sandip Mal[b]
[a] Research Scholar, School of Computing Science and Engineering,
VIT Bhopal University, Bhopal, India.
Assistant Professor, Department of Computer Technology,
Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India,
Email: ganeshyenurkar@gmail.com
[b] Assistant Professor, School of Computing Science and Engineering,
VIT Bhopal University, Bhopal, India. Email: sandip.mal@vitbhopal.ac.in

Abstract

Recent day's data is changing at an unprecedented rate in the world of data that will affect on our way to live. Challenges of big data addressing for the capturing, managing, analyzing, storing, and visualizing the big data. By these features, one can imagine the capability of big data in today's life – but certain questions may arise in future that how it will be capable? Because of the increase in data day by day and mostly advances in analytics technology. Simultaneously we also want to improve our analytical computing in terms of performance evaluation and optimization of the QOS parameter for instant processing. For research and industry, in coming years, the ability to leverage Big Data is critically increasing. In data market, data becomes an important strategic asset for survival of most of the industries in the data market. Only these industries are in the race and for those, who ignores the revolution risk are left behind and will not be able compete in the data market. The objectives of this research focus on the optimization of the QOS parameters such as accuracy, load, speed, security, trustworthiness of data by using the greedy approach of Artificial Intelligence and Machine Learning. This study comprises numerous categories of optimizing algorithms, which are referred and compared with resulting parameters to reach the specified goal. The optimized outcomes will help to design the resultant algorithm that will be capable to process any real-time data instantly. To improve the big data performance, good analysis is supported by machine learning methods. Hadoop simulator like YARN Scheduler Load Simulator (SLS) is used to solve such kind of task or problems.

Keywords: *big data, optimization model, data analysis algorithms, qos parameters*

Introduction

Nowadays, the development of big data systems continues rapidly and has also gained an unquestionable accomplishment in recent years and over the next decade. Many service areas including industry and social platform such as socialistic, internet service provider and electronic commerce as well as a variety of experimental investigation areas such as enforcement, climatology, Bionomics and structure simulations of physics and big data systems are cover by these.

Practically, tremendous, big data are represented by in testimony size and acceleration, unlikeness and variety of distinct data types and requirements of structure data processing. In the context of big data processing within an acceptable elapsed time requires advance big systems to bag, stock, investigate, and inspect them. In the fast progression of big data systems observed because of occurrence and rise

of new challenges, its complexity and diversity to study their cost effectiveness, energy productivity and performance [11] [14].

The processing of any real-life work required to minimize steps of operation leads to optimization as shown in the above fig.1. Generally, optimization applies on the known types of input data which is fed to the optimizer i.e. gradient descent algorithm. Optimizer reduces the number of steps in each operation, which is then input to the variant classes of machine learning algorithms and results into trained data. The prediction is made on the output of the machine learning model. If predicted output does not meet the required criteria, it gets input to the error function or cost function. Error function reduces the errors in the training data. This process continues till the required optimized result is obtained.

Significantly, the sources of quality and quantity of data collection are increasing drastically. So as to handle the big data, on-hand traditional algorithms are not only the computing resources that are enough but also the deployments of new processing services onto clouds are now becoming a trend in innovations. For Big data processing regardless of QoS (quality of service), an allocation of specific cloud resource approach is required. Currently, the overall QoS demand of big data becomes more challenging to incorporate with cloud while the total yield of the system is reduced [13].

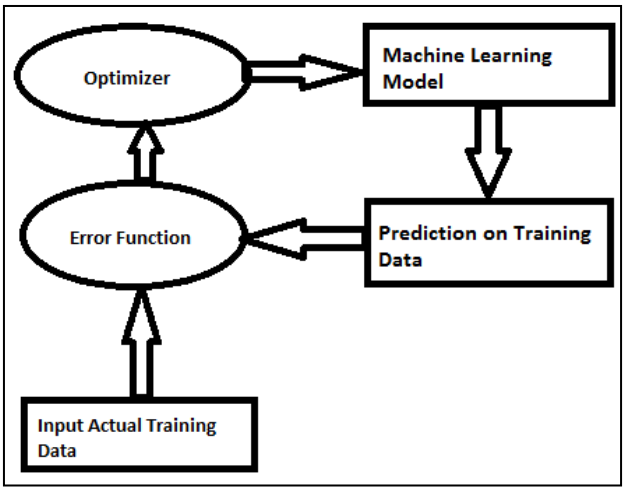

Figure 1. Machine Learning (Supervised) Optimization Model

Research Method

To cover a broad range ability of academic and scientific disciplines to process effectively with massive dataset has now become integral part of information technology world. In many contexts the big data is appearing in various form in today's world and it ranges from social media, meteorology, genomics, biological, environmental research, structure physics reproduction, commerce, and trades to health protection. Per year companies are generating more than 1000 Exabyte, within ten years it is expected to increase 20-fold. Machines and devices, business management, cloud-based solutions etc can be used to develop data. It becomes inaccessible of essential data to its users, because of loosely structured data and often incomplete. technology solutions are used to optimize QoS performance. Modern tools are used to find visualize big data, analyze and transform it. This is used to make it effective for decision making and operations [2].

The motivation from the various researchers as they introduced about their work on the various real-time applications. The proposed system will optimize the QoS performance parameters using big data-based data mining algorithms and the AI. Xindong Wu et al. [1] had recommended a model of big data that dealed with the perspective data mining by presenting HACE algorithm. It became on demand for such data models used to analyze the various information sources and interested user including modeling, security and privacy considerations. The machine learning and Hadoop based

predictive analysis model for diabetic patient was introduced by Gauri D. Kalyankar et al. [2]. They found missing values and discovered new pattern by implementing Hadoop based machine learning model by using Pima Indian diabetes data set. Predicted model was able to provide the treatment to patients based on the risk level.

Chu-Hsing Lin et al. performed the Taichung city data on land price data in through clustering algorithms. The results were visualized on google map by aggregation with R language of Hadoop HDFS and Map Reduce. Nine computing nodes with 3.5 times of acceleration were used to solve big data memory issues through cloud [3]. Siddharth Sahay et al. [4] proposed a new method in which Map Reduce framework for Hadoop Distributed File System (HDFS) is used to decrease the database scan. Cloud computing coupling processed remotely for scanning full table using Pincer-Search Algorithm. S. Suguna et al. [5] discussed the importance of e-commerce world to analyze log file for learning the user behavior system. The need of parallel transformation and stable data storage system required for analyzing web log files in large numbers. For that Hadoop framework was best suited that provided a sufficient storage. The storage is increased by distributing file System and parallel processing system for large database. The Analysis of Indian election strategies was created by Gagandeep Jagdev et.al [6]. To fulfill the objective of the research, apache Hadoop framework was used, which would comprise mining and extraction from the database and created from different districts in Punjab for election of contested MP by using thirteen information attributes which were related to different candidates. As based on previous track record of politician to set who would vote for to get correct governance? Structure clustering of huge data set using data mining style was discovered by the Kaustubh S. Chaturbhuj et al. [7] had found better initial cluster using PSO through centroids and K-means.

Maedeh Afzali et al. designed the Hadoop-MapReduce framework using Apriori algorithm [8]. Rui Han and Xiaoyi Lu summarized the lessons in which they had learned and proposed a key challenge in two aspects: firstly, to developed data capable generators for preserving the four-V's of big data in knowledge generation properties and secondly to automatic generation of benchmarking tool to test the typical application's diversity of scenario just for different system implementations and software stacks support [9] [10] [11]. The Reliability and duration of congestion metrics were measured for the first time in 2016 further utilizing the benefits of commercial speed data [12].

Mohammad Mehedi Hassan et al. [13] discovered a dynamic method for handling of provisioning to cloud resource of QoS big data processing. The various cost reducing challenging issues were addressed for big data which would incorporate the QoS demand comprehensively. A direct relation with cost minimization proved and modified efficient metric with added threshold value for Min-Min heuristic algorithm. Efficient approach in static and dynamic workload environment verified with results. Amir Daneshm et al. proposed a highly structural deterministic decomposition algorithm to reduce the sum of a possibly non-convex differentiable function and a possibly non-smooth reparable convex function [11] [14].

A.K. Reshmy et al. had overcome the problem of the unstructured dataset which was enormously huge by proposing the K-Nearest Neighbor Algorithm of machine learning which significantly improved in outcomes than the other conveniently methods used. Also, in addition networked buying system with Hadoop Map-Reduce based was studied by them [15]. Chen Hongyan et al. demanded an efficient algorithm for sorting of large dataset. The algorithm was equivalent to sort keys of task driven technology and implementation was based on windows platform. The concurrency of database support was improved by comprehensive utilization of these technologies, so that the theoretical limit value closed to performance of big data ranking. The average Limit storage used at that time for processing of compressed utilization was 46%. To analyze this problem, Nagendla Ramakrishna et al. [17] proposed a technique that was able to increase parallelism in map function skew and structure serial squeezing for informant loading. 33% data load was reduced by developed HDFS module by

overlapping compression with data loading. The allocation of tasks based on uncompressed informant size was reduced by 66% through Map-Reduce module which would compare to 15% of standard compressed information processing to allow users for achieving both performance and storage savings.

Big data-based application algorithms on classification, clustering and association were studied by Vinothini and Dr. S. Baghavathi priya [18]. In this different features and applications related to big data are studied. It also details with challenges involve in big data applications. The big data importance might evolve true value, derived from it. Mingxing Duan et al. [19] proposed a big data-based SELM algorithm for Spark structure framework to trigger the whole computing transform. The H-PMC method, Û-PMD method and method algorithm, was three parts of SLEM algorithm. Their experimental results showed that SELM algorithm obtained the highest speedup while guaranteeing the under the condition of the same parameters. The review examined by Tiago M. Fernandez proposed significant state-of-art scenarios of block chain technologies for applications in BIoT fields like smart cities, logistics, Healthcare, and energy management etc. [20]. Athira Unnikrishnan et al. compared different supervised classifiers. The required data was extracted from twitter API. Accuracy comparison among True Positive, True Negative, False Positive and False Negative was done by them. The allocation accuracies of supervised classifiers like Decision Tree, Support Vector Machine (SVM), Naive Bayes, Neural Network and k-nearest neighbor was compared on the twitter dataset. As per their research SVM allocator had highest classification accuracy than that of Nave Bayesian allocator [21].

2.1 Nowel Applications Functionalities

Today the varieties of applications in big data are increased with number of distinct technologies. Table-1 below compares some of the novel research-based applications. This includes purpose of research, real time working dataset, their performance in terms of accuracy, which is proportional to impact on the desired system with its outcomes. Major findings in our study are that there are number of real-world applications which required optimizing in terms of their performance perimeters.

Table 1*: Literature summary of some novel application*

Ref. No.	Objective	Dataset Used	Methodology Used	% Accuracy	Impact	Major Improvement needed (Yes/No)
[22]	To derive sequential pattern mining for Ensemble learning based on deep learning model.	Laboratory Dataset	Deep learning models -Long Short-Term Memory (LSTM) -Hopfield.	97.17 %.	Strong	No
[23]	To study credit driven assessment framework for SMEs including big data from business, government, social media & networks.	Small SMEs credit rating data	Financial and non-financial data	76.97%	Slight	Yes
[24]	To detect traffic congestion in Indonesia, especially DKI Jakarta	Traffic Data	Classification methods such as SVM	98.92%	Slight	Yes

	planted on Twitter Data adopting Machine Learning.		Linear , k-NN and Naive Bayes			
[25]	To detect falls of elderly people in indoor environments through IoT.	SisFall dataset	IoT, ML processing techniques based on decision trees	91.67%	Strong	Yes
[25]	To detect falls of elderly people in indoor environments through IoT.	SisFall dataset	IoT, ML processing techniques based on decision trees	91.67%	Strong	Yes
[27]	To study graph embedding and statistical relational learning method for retrieval of socio-economic data	Fiscal Dataset	Stacked method with SRL framework	88%	Strong	Yes
[28]	To provide advice for banks, a multi-class learning was studied from label proportions, and apply it to better management of customer relationships	Customer Dataset was selected with customer ID, age, and credit score etc., attributes.	EML LLP-EML	80%	Slight	Yes
[29]	To study semi-supervised learning as a method for reducing the effort and time utilized in data labeling.	IMDB dataset	Supervised deep neural network, semi s upervised deep neural network	882%	Slight	Yes
[30]	To check the impact of BDQM in any big data project.	SA Data	LDA, SVM, NN	Disastrou s SA accuracy (32.40%).	Slight	Yes
[31]	To study the big data technologies for real-time occupancy detection	Occupancy Detection Dataset	IoT, Storm and Kaa	95%	Strong	Yes
[32]	To Investigate the impact of using both touch screen-based and sensor-based features in an authentication model using deep learning methods.	HMOG dataset	Training network on HMOG	88%	Strong	Yes

The purpose of computation of units in graphics processing are used for speeding up [33] [44]. A CPU cluster over incomputable datasets via traditional parallel approaches. Image and video processing used for satellites, telescope, recognition of biological and medical images [34] [40], computer vision [39], automatic picture or audio annotations [34] [38]. Evdokia Kassela et al. developed a big data based analytical subsystem which gives resourceful solution. The solution is obtained using data analytics and machine learning technologies for elastic energy and base stations [35]. Sheila Alemany et al. designed a model that predicted the Unisys Weather data to collect the time frequency [36]. Dinithi Nallaperuma et al. [37] designed a new smart and dynamic traffic to analyze and integrate the AI element for problem-solving time in traffic management platform to control captured data streams. Recently rise of AI accelerators and cryptographic to optimized Intel's Neural Processors [39] [40] for new hardware architecture generation has evolved.

Alberto Cano et al. [41] did survey and proposed to analyze current trends of GPU computing for large-scale data mining and discussed GPU architecture for handling volume and velocity of big data, they also identified limitation factors for scalability of the problems open issues and future directions. In context of big data and machine learning the Classical methodologies in data mining could not be able to process massive and high-speed volumes of information. J. Alcala-Fdeza et al. [42] presented new aspects of KEEL dataset which included the partition of it and have shown results. The idea of numerical precision training and inference were put to enabling deep learning inference and training of lower numerical precision by Andres Rodriguez et al. [43]. They had discovered that how deep learning frameworks were taking more advantageous of these reduced lower numerical precision functions between different categories of numerical precisions. A Machine Learning and Deep Learning frameworks and libraries survey of for large-scale data mining was done by Giang Nguyen et al. Data mining project tools used to decide and select challenging frameworks among divergent Machine Learning and Deep Learning user community's different applicable areas like libraries, tools, and approaches [44].

The commonly used machine learning algorithms results were applied to the clinical data automatic classifier to diagnosed different diseases. KNN and SVM played an essential role for their purposed system [45]. To detect the process of network storage steganography of machine learning algorithm were proposed by Cho D.X. et al. [46]. Riad Akrour et al. proposed a new optimization trajectory-based policy for dynamic algorithm linearization [47][48].

Zeyi Wen et al. presented a powerful Thunder SVM open-source software toolkit which feat Graphics Processing Units (GPUs) at high performance and multi-core CPUs. Multiple language interfaces including massive parallel algorithm to large-scale data extends for a fraction of cost with a traditional high-performance C/C++, Python, R and MATLAB were used by thunder [48]. Moritz Hardt et al. proved that the stochastic gradient descent global optimizer effectively converges the peak likelihood purpose of an unknown linear time-inherent dynamical system generated by the system in their research from a sequence of noisy observations [49].

An adjacent pair of point dispute relationship along with interaction investigation exploited through Gian-Andrea Thanei et al. [50]. They discovered a run time algorithm for interaction search that was sub quadratic under strong assumption to search linearity for very strong interactions. An approach of detecting fake data source, which complements the piece-by-piece fake data detection algorithms in previous studies were proposed by Xiaofan Li et al. [51]. Assorted paths suggested figuring out the statistics piece by piece by them to understand a source by pushing data helpfully. Andrew Cotter et al. [52] explains some real-world issues to resolve most comprehensive study to-date of training classifiers with a extensive array of rate constraints. It mostly consist of new theoretical, algorithmic, and experimental results along with practical insights and guidance for using rate. Christoph D. Hofer et al. studied the approach to learned representation of task-specific barcodes. On other hand, their aim was to adapt the learning problem by preserving theoretical properties such as stability. This

projected a vector space barcode into a unite dimensional adopting a set of parameterized functional, so called design aspects for which they provided a universal construction scheme [53].
Jaouad Mourtada et al. [54] carried out the concern investigation of the standard exponential weights (Hedge) methods in the probability expert setting, closure of a gap in the existing research. A distributed optimization framework was proposed by Can Karakus et al. where they encoded and completed built-in redundancy dataset representation and the dynamically straggling nodes were employed as removed or as "erasures" at every repetition compensated by the embedded repetition. They appliance the recommended technique on Amazon EC2 clusters [55]. Akshay Krishnamurthy et al. presented a new active learning method used a square loss oracle to investigate the form space and drive the query planning for cost-sensitive multiclass classification that guarantees on running time, generalization error and label complexity. The main algorithmic research was a new way to figure out the higher and lower costs predicted by a regression function in the form space [56].
A novel sketching method robust frequent direction (RFD) was proposed by Cheng Chen et al., achieved a better performance than baselines by using second order online learning algorithms [55][57]. A labeled noise of multiclass classification with mixed models of membership and partial classification labels were examined by Julian Katz-Samuels et al. [58] characterized the three machine learning problems through mutual contamination models. Ben Dai et al. discussed a smooth collaborative system of recommendations which integrates the network structure of user-item pairs to improve prediction accuracy which would provide a flexible framework to exploit the covariate information, such as user demographics, item contents, and social network information for users and/or items. Although the proposed method as formulated was based on the latent factor model [59].
The forecasting approaches of building's appliances for load development were integrated to achieve occupancy prediction and control of context-driven presented by S. Hadri et al. [60]. The energy consumption was forecasted by using IoT-Big data-based platform. They also suggested the recorded predictive data model for ARIMA, Random Forest (RF). Sarima et al. introduced a novel approach for sequential deep learning model to enhance the efficiency of accuracy during recognition of pattern mining. To enhance the ensemble learning sequential patterns scheme of generalization for mining algorithms [61]. Fatma Chiheb et al. [62] provided big data analytics integrated theoretical model that process the decision-making processes into various phases. To improve the organizations decision quality through theoretical DMP-BDE Model was used to allow decision-making based phases of Big Data processes. Algorithmic optimization of a research problem leads to the major role in the problem solving in terms of time and space complexity. Such a real-world problem in in data analysis required the optimization for their faster processing of data.

2.2 Optimization Algorithms and their Research Scope

Following table-2 shows that the existing optimization techniques in big data analysis. Table-2 shows that the algorithm creation, dataset used, future research scope and their purpose of use in real world problem solving Big data is radically changing in biomedical field to data analysis and interpretation with major challenges by advanced approaches. A powerful and effective system in personalized medicine with significant scientific and technical developments was created by Davide Cirillo et al. used a Biomedical Big Data [63]. The big data analysis association between outcomes of reported patients, Observer those who had reported toxicities and the radiation therapy in aspect of life of head and neck patients cured with cancer was proposed by Joel R. Wilkie et al. [64], their studies identified that the lifestyle, activities, and fatigue become the considerable items in clinical Questionnaire.
Zachary N. Harris et al. proposed two approaches, first one was profiling read-based Approach and secondly to reduced assembly dataset-based approach to analyze efficiently on data set with large-scale. Multiple machine learning techniques was to predict unknown samples precisely incorporated in the pipeline [65]. A multi-scale region positioned convolutional neural network (MR-CNN) with

small traffic sign perception was presented by Juan Du et al. Their experimental results were too superior for detecting small traffic sign and also achieved a significant performance [66][67]. Their WASS R-CNN method was achieved detection accuracy on the impressive challenging PASCAL VOC 2007 dataset.

The problem of affective image emotion recognition analysis to its complexity and subjectivity was investigated by Tianrong Rao et al. their framework was automatically detecting an emotional region in multilevel deep feature maps [68].

The utilization method of big data technology explores in scientific research of military data administration and the development of scientific research big data platform. A scientific research technical platform was built to provide big data military enterprises application on platform [69]. The reviewed cleansing process of big data challenge for data cleansing and methods for data cleansing were available. A value and veracity of the data must be taken into consideration for the evaluation of proposed methods [70]. Data analytics discovered the predictive power behind collected data addressed by using cleaning criteria. To understand knowledge discovery (KDD) in database, extracting information patterns. The presented framework accessed the distinct KDD processes starting from scratch [71].

The high frequency big data analytical frameworks studied by combining high effective learning algorithms with map reduce computation using PANFIS [72]. The high-performance distributed computing databases and architecture were studied for multiple exploitation re-configurable computing specific processing application like CPUs, and FPGAs efficiently; the data quality can be improved of outlier's detection, data cleaning and data of missing interpolation [73]. A new patent segmentation method and patent analysis simplification was described by Maryam Habibia et al.

The method was builds upon unsupervised text segmentation, text tiling and applied CRF sequential classifiers for segment classification [74]. Salah Ud Dina Junming Shao et al. proposed an online semi-supervised learning algorithm with a set of micro- clusters by modeling concept drifts method for capturing effective data stream learning. A quality based optimal pricing model of big data market was proposed by Jian Yang et al. [76] which allowed to maximize optimized quality level data platform owners and profits subscription fees.

Table 2: ***Shows Research Scope and Purpose of Optimization Algorithms***

Ref.	Algorithm	Research Scope	Purpose of Use	Dataset Used
[47]	MOTO	Completed	Optimization of existing linearization dynamics algorithm	---
[49]	SGD	Open	Training recurrent neural networks	---
[51]	Developed theoretic model to capture the source of information	Open	Detection of Fake Data Source	Model data traffic
[52]	"proxy-Lagrangian" formulation	Open	Optimized applications with Non-Differential Constraints	Bank Marketing UCI benchmark, Adult income UCI, ProPublica's COMPAS recidivism
[53]	SVM	Open	To learned barcodes with a task-specific representation	EEG, MPEG-7, Animal 2D shape

[54]	Hedge, SGD	Open	Optimization of standard exponential weights (Hedge)	---
[55]	Gradient Descent, L-BFGS, and Proximal Gradient	Open	Optimization of Distributed Framework	Amazon EC2 clusters
[56]	COAL	Completed	Active Learning for Cost-Sensitive Classification	ImageNet40 and RCV1-v2
[57]	Hyper parameter free online Newton algorithm	Open	To optimize online learning algorithm with RFD	a9a, gisette, sido0, farm-ads, rcv1 and real-sim
[58]	Membership with Multiclass Classification models, and with partial labels classification	Open	Decontamination of Mutual Contamination Models	MNIST, Iris and Breast Cancer Wisconsin
[59]	Divide-and-Conquer	Open	smooth collaborative recommender system	online music dataset from the Last.fm

2.3 Applications Comparison

Following table-3 shows that some of the notable big data analytical application having parameters likes data types, methods used, accuracy of an application, no. of sample data tested to achieve the specified results and their research focus on area.

Table 3: ***Notable applications in Machine Learning***

Task	**Method**	**% Accuracy**	**No. of Sample Tested**	**Research Focus**	**Ref**
Big Data Analytics for health care and Personalized Medicine	GLM	99.90	10 Petabytes	Healthcare and Biomedical Research	[63]
To access associations between varieties of patient-reported outcomes (PROs)	ORT, QOL	78	612	Clinical Research (Questionnaires in the clinic)	[64]
Analysis of metagenomic data.	Read based and Assembly based approach and Random Forest Prediction	66.66, 60 and 89.7 Respectively	30	General	[65]
Small Traffic Sign Recognition	MR-CNN	71.3	1000	General	[66]
Object Detection	WASS R-CNN	85	5011	General	[67]

Image Classification	Emotion extraction for Multi-Level Deep Representation	R-CNN	87.51	100	General [68]

Methodology

Results of the research are explained in this section. Along with the result part, comprehensive discussions are also described in this section. Figures, graphs, tables are used to analyze the results, this will make the reader understand easily [2, 5]. The discussion can be made in several sub-chapters.

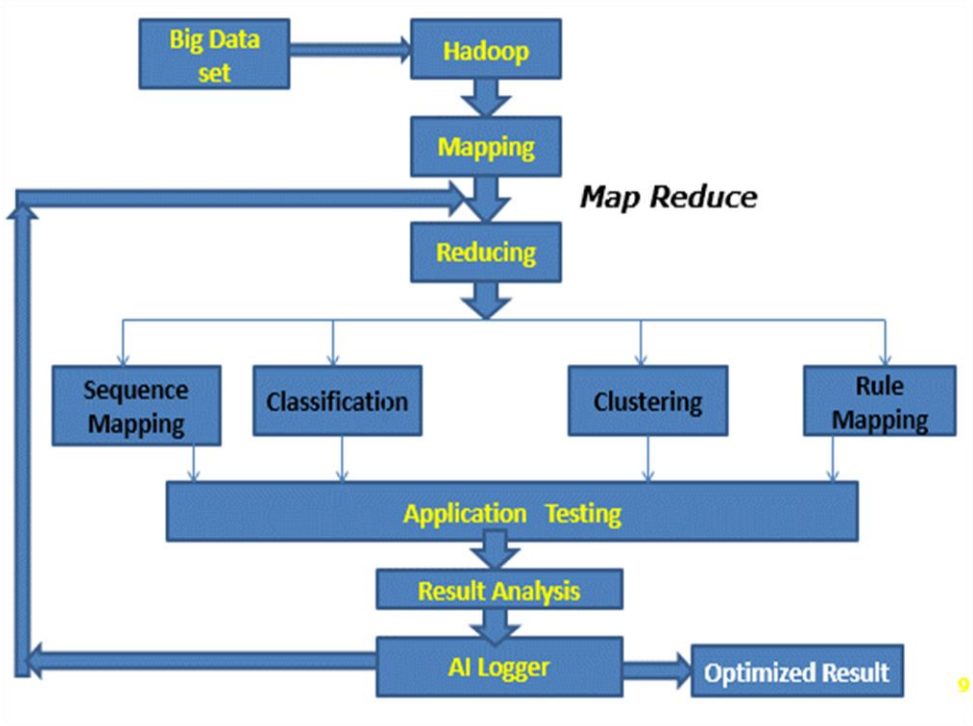

Figure 2: Proposed System Workflow

The research plan comprises as follow: -

a) Data
b) Analysis Techniques
c) Applying advanced data mining techniques.
d) Use of Machine learning techniques
e) Applicability
f) Advanced testing through AI for optimized result (By newly adopted Algorithm)

Discussion

The data set and methods presented here is to identify an appropriate method of analysis. It is worth discussing these interesting facts revealed by the broad categories of research application in above discussion. Here we compared the results of the existing approaches with those of the traditional one.

The limitations of the existing studies naturally exclude some sort of optimization. There were also some important differences and key findings that will act as pitfalls investigation problem in our research. However, it is possible to optimize or enhance the QoS parameter through advanced

machine learning algorithm with spark framework. Nonetheless, we believe that it is well justified in the above table to carry out our research.

The first approach in the monotonic improvement of MOTO algorithm [47] with more refined extension is to bias-variance trade-off. In [61], the scope of enhancement of sequential patterns mining algorithms in ensemble learning model is introduced. There may be a scope as reviewed in [64], that investigation of PRO to improve outcomes in other body parts also. Possibility to increase the prediction accuracy of massive Meta genomic data using abundance-based machine learning is suggested in [65].

Conclusion

In summary, the reviewed literature suggests that the systematic study of the different approaches related to optimization techniques towards Big data analytics through variety of machine learning algorithms. Researchers had done the specific work of their application and optimized only for the dedicated application. Needs to enhance the capability of the machine learning model to optimize the certain Quality of Service parameters. In future, our aim is to develop a universal algorithmic system to optimize the Quality-of-Service parameters by using the combination of the premier technologies like Big Database Mining Algorithm using Machine Learning and Artificial Intelligence by considering real world application.So that the upcoming system can be able to improve considerable performance of the big data analytics.

References

[1] Xindong Wu , Xingquan Zhu , Gong-Qing Wu , Wei Ding, (2014). Data mining with big data. IEEE 10.1109/TKDE.2013.109 Transactions on Knowledge and Data Engineering (Volume: 26, Issue: 1).

[2] Gauri D. Kalyankar , Shivananda R. Poojara ,Nagaraj V. Dharwadkar, (2017). Predictive analysis of diabetic patient data using machine learning and Hadoop. 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC),DOI: 10.1109/I-SMAC.2017.8058253.

[3] Chu-Hsing Lin , Jung-Chun Liu , Tsung-Chi Peng (2017). Performance evaluation of cluster algorithms for Big Data analysis on cloud. International Conference on Applied System Innovation (ICASI).

[4] Siddharth Sahay , Suruchi Khetarpal , Tribikram Pradhan (2016). Hybrid data mining algorithm in cloud computing using MapReduce framework. International Conference on Advanced Communication Control and Computing Technologies (ICACCCT),

[5] S. Suguna , M. Vithya , J. I. Christy Eunaicy (2016). Big data analysis in e-commerce system using Hadoop MapReduce. International Conference on Inventive Computation Technologies (ICICT), 26-27.

[6] Gagandeep Jagdev ; Amandeep Kaur (2016). Analyzing and scripting indian election strategies using big data via Apache Hadoop framework. 5th International Conference on Wireless Networks and Embedded Systems (WECON).

[7] Kaustubh S. Chaturbhuj , Gauri Chaudhary (2016). Parallel clustering of large data set on Hadoop using data mining techniques. World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), 29 Feb.-1.

[8] Maedeh Afzali , Nishant Singh (2016). Hadoop-MapReduce: A platform for mining large datasets. 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 16-18.

[9] Sudhakar Singh, Rakhi Garg, P.K. Mishra (2017). Performance optimization of MapReduce-

based Apriori algorithm on Hadoop cluster. Computers and Electrical Engineering 000(2017)1-17, www.elsevier.com/locate/compeleceng.

[10] M. Omair Shafiq,Maryam Fekri, Rami Ibrahim (2017). MapReduce Based Classification for Fault Detection in Big Data Applications. 16th IEEE International Conference on Machine Learning and Applications,

[11] Rui Han and , Xiaoyi Lu. On Big Data Benchmarking.

[12] "Bigdata performance metrics", 2016 LOS MONITORING REPORT - Prepared by Iteris, Inc. page no. 52-61.

[13] Mohammad Mehedi Hassan, Biao Song, M. Shamim Hossain and Atif Alamri (2014). QoS-aware Resource Provisioning for Big Data Processing in Cloud Computing Enviornment. International Conference on Computational Science and Computational Intelligence, 978-1-4799-3010-4/14 $31.00 © 2014 IEEE.

[14] Amir Daneshmand, Francisco Facchinei, Vyacheslav Kungurtsev, and Gesualdo Scutari (2014). Flexible Selective Parallel Algorithms for Big Data Optimization", 178-1-4799-8297-4/14/$31.00©2014IEEE.

[15] A.K. Reshmy, D. Paulraj (2015). An Efficient Unstructured Big Data Analysis Method for Enhancing Performance using Machine Learning Algorithm. International Conference on Circuit, Power and Computing Technologies (ICCPCT), 978-1-4799-7075-9/15/$31.00 ©2015 IEEE.

[16] Chen Hongyan, Wan Junwei, Lu Xianli (2017). Research and Implementation of Database High Performance Sorting Algorithm with Big Data. IEEE 2nd International Conference on Big Data Analysis.

[17] Nagendla Ramakrishnaiah and Sirigiri Konda Reddy (2017). Performance Analysis of Matrix and Graph Computations using Data Compression Techniques in MPI and Hadoop MapReduce in Big Data Framework. IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Veltech Dr.RR & Dr.SR University, Chennai, T.N., India. pp.54-62.

[18] Vinothini and Dr. S. Baghavathi priya (2017). Survey of Machine Learning Methods for Big DataApplications. International Conference on Computational Intelligence in Data Science(ICCIDS)

[19] Mingxing Duan, Kenli Li, Xiangke Liao and Keqin Li (2018). A Parallel Multiclassification Algorithm for Big Data Using an Extreme Learning Machine. IEEE Transactions On Neural Networks And Learning Systems, Vol. 29, No. 6.

[20] Tiago M. Fernández-Caramés and Paula Fraga-Lamas (2018). A Review on the Use of Blockchain for the Internet of Things. Received April 11, 2018, accepted May 29, 2018, date of publication May 31, 2018, date of current version July 6, 2018. Digital Object Identifier 10.1109/ACCESS.2018.2842685.

[21] Athira Unnikrishnan Uma Narayanan Dr. Shelbi Joseph (2017). Performance Analysis of Various Supervised Algorithms on Big Data. International Conference on Energy.

[22] Choukri Djellali, Mehdi adda (2019). A new Deep learning model for Sequential Pattern Mining using Ensemble learning and Models selection Taking Mobile Activity Recognition as a case. The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC) August 19-21, 2019, Halifax, Canada, Procedia Computer Science 155 (2019) 129–136

[23] LIU Yadi, , SONG Yuning, YU Jiayue, XIE Yingfa, WANG Yiyuan, ZENG Xiaoping (2019). Big-data-driven Model Construction and Empirical Analysis of SMEs Credit Assessment in China. International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2018 , Procedia Computer Science 147 (2019) 613–619

[24] Muhammad Taufiq Zulfikar, Suharjito (2019). Detection Traffic Congestion Based on Twitter Data using Machine Learning. 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12–13 September 2019 , Procedia Computer Science 157 (2019) 118–124

[25] Diana Yacchiremaa,b, Jara Suárez de Puga, Carlos Palau, Manuel Esteve (2018). Fall detection system for elderly people using IoT and Big Data. The 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018) , Procedia Computer Science 130 (2018) 603–610.

[26] Hamid R. Darabi, Daniel Tsinis, Kevin Zecchini, Winthrop F.Whitcomb, Alexander Liss (2018). Forecasting Mortality Risk for Patients Admitted to Intensive Care Units Using Machine Learning. Complex Adaptive Systems Conference with Theme: Cyber Physical Systems and Deep Learning, CAS 2018, 5 November – 7 November 2018, Chicago, Communication, Data Analytics and Soft Computing (ICECDS-2017) Illinois, USA , Procedia Computer Science 140 (2018) 306–313.

[27] Alexander Kalinina, Danila Vaganova, Klavdiya Bocheninaa(2019). Improving statistical relational learning with graph embeddings for socio-economic data retrieval. 8th International Young Scientist Conference on Computational Science , Procedia Computer Science 156 (2019) 235–244

[28] Yaxing Qian, Qiang Tong, Bo Wang (2019). Multi-Class Learning from Label Proportions for Bank Customer Classification. 7th International Conference on Information Technology and Quantitative Management (ITQM 2019) , Procedia Computer Science 162 (2019) 421–428.

[29] Vivian Lay Shan Lee, Keng Hoon Gan, Tien Ping Tan, Rosni Abdullah (2019). Semi-supervised Learning for Sentiment Classification using Small Number of Labeled Data. The Fifth Information Systems International Conference 2019, Procedia Computer Science 161 (2019) 577–584.

[30] Imane El Alaoui, and Youssef Gahi (2019). The Impact of Big Data Quality on Sentiment Analysis Approaches. International Workshop on Emerging Networks and Communications (IWENC 2019) November 4-7, 2019, Coimbra, Portugal, Procedia Computer Science 160 (2019) 803–810.

[31] H. Elkhoukhi, Y. NaitMalek, A. Berouine, M. Bakhouya, D. Elouadghiri, A. Berouine (2018). Towards a Real-time Occupancy Detection Approach for Smart Buildings. The 15th International Conference on Mobile Systems and Pervasive Computing, Procedia Computer Science 134 (2018) 114–120.

[32] Hasan Can Volaka, Gulfem Alptekin, Okan Engin Basar, Mustafa Isbilen, Ozlem Durmaz Incel (2019). Towards Continuous Authentication on Mobile Phones using Deep Learning Models. The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC) August 19-21, 2019, Halifax, Canada , Procedia Computer Science 155 (2019) 177–184.

[33] Abdiansah A, Wardoyo R (2015). Time complexity analysis of support vector machines (SVM) in LibSVM. Int J Comput Appl 128(3):28–34Google Scholar.

[34] Akiba T (2017). Performance of distributed deep learning using ChainerMN", https://chainer.org/general/2017/02/08/Performance-of-Distributed-Deep-Learning-Using hainerMN.html.Accessed.

[35] Evdokia Kassela, Nikodimos Provatas (2019). BigOptiBase: Big Data Analytics for Base StationEnergy Consumption Optimization. IEEE International Conference on Big Data (Big Data), 978-1-7281-0858-2/19/$31.00 ©2019 IEEE

[36] Sheila Alemany , Jonathan Beltran (2018). Predicting hurricane trajectories using a recurrent neural network. arXiv:1802.02548v3 [cs.LG].

[37] Dinithi Nallaperuma, Rashmika Nawaratne (2019). Online Incremental Machine Learning

Platform for Big Data-Driven Smart Traffic Management. IEEE Transaction on Intelligent Transportation Systems, VOL. 20, NO. 12.

[38] Anaconda (2018) Anaconda (2018). the most popular python data science platfor. https://www.anaconda.com/what-is-anaconda/ . Accessed 20 Oct 2018.

[39] AnacondaCloudera (2016) Anaconda for Cloudera (2018). data science with python made easy for big data. http://know.continuum.io/anaconda-for-cloudera.html. Accessed 20 Oct 2018.

[40] Andres R et al (2018). Lower numerical precision deep learning inference and training Intel. https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training. Accessed 18 Sept 2018

[41] Alberto Cano (2017). A survey on graphic processing unit computing for large-scale data mining. WIREs Data Mining Knowl Discov 2017, e1232.

[42] Alcala-Fdez J, Fernandez A, Luengo J, Derrac J, Garcia S, Sanchez L, Herrera F (2011). KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. J. of Mult.-Valued Logic & Soft Computing, Vol. 17, pp. 255–287, ©2011 Old City Publishing, Inc. Published by license under the OCP Science imprint, a member of the Old City Publishing Group.

[43] Andres Rodriguez, Eden Segal, Etay Meiri, Evarist Fomenko, Young Jim Kim, Haihao Shen, and Barukh Ziv (2018). Lower Numerical Precision Deep Learning Inference and Training. Intel White paper.

[44] Giang Nguyen1, Stefan Dlugolinsky1, Martin Bobák1, Viet Tran1, Álvaro López García2, Ignacio Heredia2, Peter Malík1 , Ladislav Hluchy (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review (2019) 52:77–124 (Springer), https://doi.org/10.1007/s10462-018-09679-z Published online: 19 January 2019 © The Author(s) 2019.

[45] Giorgio Biagettia, Paolo Crippaa,_, Laura Falaschettia, Giulia Tanonia, Claudio Turchettia (2018). A comparative study of machine learning algorithms for physiological signal classification. Procedia Computer Science 126 (2018) 1977–1984.

[46] Cho D.X, Thuong D.T. H,Dung N.K. (2019). A method of detecting storage based network steganography using machine learning. 8th International Congress of Information and Communication Technology, ICICT 2019,ScienceDirect Procedia Computer Science 154 (2019) 543–548

[47] Riad Akrour, Abbas Abdulsalami, Hany Abdulsamad, Jan Peter, Gerhard Neumann (2018). Model-Free Trajectory-based Policy Optimization with Monotonic Improvement. Journal of Machine Learning Research 19 (2018) 1-25 Submitted 6/17; Revised 4/18; Published 8/18.

[48] Zeyi Wen, Jiashuai Shi, Qinbin Li, Bingsheng He, Jian Chen (2018). ThunderSVM: A Fast SVM Library on GPUs and CPUs. Journal of Machine Learning Research 19 (2018) 1-5 Submitted 12/17; Revised 6/18; Published 7/18.

[49] Moritz Hardt, Tengyu Ma, Benjamin Recht (2018). Gradient Descent Learns Linear Dynamical Systems. Journal of Machine Learning Research 19, 1-44 Submitted 9/16; Revised 7/18; Published 8/18.

[50] Gian-Andrea Thanei, Nicolai Meinshausen, Rajen D. Shah (2018). The xyz algorithm for fast interaction search in high-dimensional data. Journal of Machine Learning Research 19,1-42 Submitted 10/16; Revised 12/17; Published 8/18.

[51] Xiaofan Li, Andrew B. Whinston (2020). A model of fake data in data-driven analysis. Journal of Machine Learning Research 21, 1-26 Submitted 6/17; Revised 7/19; Published 02/2.

[52] Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan (2019). Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. Journal of Machine Learning Research 20. 1-59 Submitted 9/18; Revised

10/19; Published 11/19.

[53] Christoph D. Hofer, Roland Kwitt, Marc Niethammer (2019). Learning Representations of Persistence Barcodes. Journal of Machine Learning Research 20. 1-45 Submitted 6/18; Revised 3/19; Published 7/19.

[54] Jaouad Mourtada, St´ephane Gaiffas (2019). On the optimality of the Hedge algorithm in the stochastic regime. Journal of Machine Learning Research 20. 1-28 Submitted 12/18; Revised 4/19; Published 5/19.

[55] Can Karakus, Yifan Sun, Suhas Diggavi, Julian Wotao Yin (2019). Redundancy Techniques for Straggler Mitigation in Distributed Optimization and Learning. Journal of Machine Learning Research 20. 1-47 Submitted 3/18; Revised 4/19; Published 4/19.

[56] Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daum´e III, John Langford (2019). Active Learning for Cost-Sensitive Classification. Journal of Machine Learning Research 20. 1-50 Submitted 11/17; Revised 3/19; Published 4/19.

[57] Luo Luo, Cheng Chen, Zhihua Zhang, Wu-Jun Li, Tong Zhang (2019). Robust Frequent Directions with Application in Online Learning. Journal of Machine Learning Research 20. 1-41 Submitted 12/17; Revised 8/18; Published 2/19.

[58] Julian Katz-Samuels, Gilles Blanchard, Clayton Scott (2019). Decontamination of Mutual Contamination Models. Journal of Machine Learning Research 20. 1-57 Submitted 9/17; Revised 11/18; Published 1/19.

[59] Ben Dai, Junhui Wang, Xiaotong Shen, Annie Qu (2019). Smooth neighborhood recommender systems", Journal of Machine Learning Research 20. 1-24 Submitted 10/17; Published 2/19.

[60] S. Hadri, Y. Naitmalek, M. Najib, M. Bakhouya, Y. Fakhri, M. Elaroussi (2019). A Comparative Study of Predictive Approaches for Load Forecasting in Smart Buildings. The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, November 4-7, Coimbra, Portugal.

[61] Choukri Djellali_, Mehdi adda (2019). A new Deep learning model for Sequential Pattern Mining using Ensemble learning and Models selection Taking Mobile Activity Recognition as a case. The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC) August 19-21, , Halifax, Canada.

[62] Fatma Chiheb, Fatima Boumahdi, Hafida Bouarfa (2019). A New Model for Integrating Big Data into Phases of Decision-Making Process. The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40) April 29 - May 2, 2019, Leuven, Belgium.

[63] Davide Cirillo, Alfonso Valencio (2019). Big Data Analytics for PersonalizedMedicine. https://doi.org/10.1016/j.copbio.2019.03.004,www.sciencedirect.com

[64] Joel R. Wilkie,Michille L. Mierzwa, John Yao,Avraham Eisbruch,Mary Feng,Grant Weyburne,Xiaoping Chen,Lynn Holevinski, Charles S, Mayo (2019). Radiotherapy and Oncology. https://doi.org/10.1016/ j.radonc. 2019.04.030, Radiotherapy and Oncology 137 (2019) 167-174.

[65] Zachary N. Harris, Eliza Dhungel, Matthew Mosior and Tae-Hyuk Ahn (2019). Massive metageonomic data analysis using abundance-based machine learning. Harris et al. Biology Direct (2019) 14:12https://doi.org/10.1186/s13062-019-0242-0.

[66] Zhigang Liu, Jaun Du, Feng Tian, JiazengWen. MR-CNN: A multi-scale region based convolutional neural network (MR-CNN) for small traffic sign recognition. http://www.ieee.org/publications_standards/ publications/tights/index.html,2169-3536 IEEE.

[67] Xing-Gang Wang, Jia-Si Wang,Wen-Yu Liu (2019). Weekely and SemiSupervised Fast region Based CNN F for Object Detection. Journal of Computer Science and Technology 34(6):1269-1278 Nov 2019, DOI 10.100/s11390-019-1975-z

[68] Tianrong Rao,Xiaoxu Li,Hiamin Zhang,Min Xu (2019). Multi-Levl Rgion Based

Convolutional Network for Image Emotion Classification. https://doi.org/10.1016/j.neucom.2018.12.053,Neurocomputing 333,429-439.

[69] Wang Kun, Liu Tong, Xie Xiaodan (2018). Application of Big Data Technology in Scientific Research Data Management of Military Enterprisz. International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2018, Procedia Computer Science 147, 556–561

[70] Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon (2019). A Review on Data Cleansing Methods for Big Data. The Fifth Information Systems International Conference , Procedia Computer Science 161 (2019) 731–738

[71] Hicham Moad Safhi, Bouchra Frikh, Brahim Ouhbi (2019). Assessing reliability of Big Data Knowledge Discovery process. Second International Conference on Intelligent Computing in Data Sciences (ICDS 2018), Procedia Computer Science 148, 30–36

[72] Choiru Za'in, Mahardhika Pratama, Edwin Lughofer, Meftahul Ferdaus, Qing Cai (2018). Big Data Analytics based on PANFIS MapReduce", INNS Conference on Big Data and Deep Learning , Procedia Computer Science 144 (2018) 140–152

[73] Olivier Debauche, Sidi Ahmed Mahmoudi, Sa¨ıd Mahmoudi, Pierre Manneback (2018). Cloud Platform using Big Data and HPC Technologies for Distributed and Parallels Treatment. The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, Procedia Computer Science 141 (2018) 112–118.

[74] Maryam Habibia, Astrid Rheinlaenderb, Wolfgang Thielemannb, Robert Adamsb, Peter Fischerc Sylvia Krolkiewiczb, David Luis, Wiegandta Ulf Leser PatSeg (2020). A Sequential Patent Segmentation Approach. Big Data Research, 100133, https://doi.org/10.1016/j.bdr.2020.1001332214-5796/©2020 Elsevier Inc.

[75] Salah Ud Dina Junming Shao, Waqar Alib Jiaming LiuaYuYe (2020). Online reliable semi-supervised learning on evolving data streams. Big Data Research 153-171, https:// doi.org/10.1016/j.ins. 2020.03.052.

[76] Jian Yang, Chongchong Zhao, and Chunxiao Xing (2019). Big Data Market Optimization Pricing Model Based on Data Quality. Hindawi Complexity Volume, Article ID 5964068, 10 pageshttps://doi.org/10.1155.\

[77] Whitcomb, J., Borko, H., & Liston D. (2009). Promising professional development models and practices. *Journal of Teacher Education*, *60*(3), 207–12.