A Comprehensive Study of Machine Learning Mechanisms On Big data

Mounica Vennapusa, Srikanth Bhyrapuneni

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

***Index Terms*: About Machine Learning, Bigdata, Deep Learning, Neural Networks.**

I. INTRODUCTION

Data is the backbone of all businesses. As a result of developments in Social Media, Mobiles, Web technologies and Sensing devices, the amount of data is increasing at an unusual rate. For example the amount of data we deliver everyday is truly exciting. There are 2.5 quintillion bytes of data created everyday at current rate. Data is increasing at a rapid pace. By 2020 the new data generated for every individual per second will be the approximate amount of 1.7 mega bytes. By 2020, the accumulated information of Bigdata will increase from 4.4 Zetta bytes to roughly 44 Zetta bytes or 44 trillion giga bytes[17]. This Bigdata owns massive increase in terms of business value in variety of fields such as medical field, financial services, health care, transportation and online advertising[2][12]. However the traditional methods are facing difficulty with this huge amount of data[1].

A. Machine Learning

Machine Learning is about learning to improve in the future later dependent on what was knowledgeable previously. The objective is to formulate learning algorithms that do the learning naturally without human assistance or supervision. Machine Learning is a sub area of artificial intelligence which empowers software applications to get into a state of self-learning without being explicitly programmed [8][18]. When presented to new data, these systems are empowered to learn, change, grow and implement by themselves. Machine Learning concentrates on the advancement of systems that can get to information and utilize it from themselves [42]. Machine Learning helps to identify the data and trends. The aim is to enable the computers to learn consequently and modify actions appropriately.

B. Bigdata

Bigdata is an emerging term that outlines any volume less measure of structure, semi-structured and unstructured data that has the ability to be mined for information. Bigdata is relatively new, the way towards gathering and storing immense measures of data for inevitable analysis[37].Bigdata is substantial datasets and the category of computing strategies and technologies that are utilized to handle these large datasets[9]. Bigdata is the data that have greater variety occurring in expanding volumes with higher velocity.

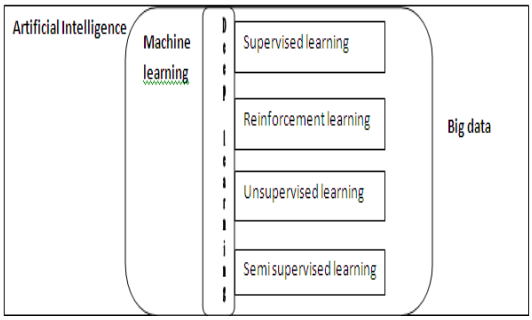

Fig 1 Machine Learning on Bigdata

II. RELATED WORK

This section widely represents the research exploration and advancements that has occurred in past years.

Lidong Wang discussed some methods in Data Mining and Machine Learning and given their comparison of advantages and disadvantages; analyzed IT challenges and service infrastructure of Big Data [1] Junfei Qiu, surveyed the latest Machine Learning techniques and introduced some of recent learning methods to solve Big Data problem; outlined the critical challenges, research trends, open issues of Machine Learning methods for Big Data processing. [3] Youming, proposed reference anatomy of Machine Learning for Big Data Analytics; analyzed research challenges and issues of Machine associated with Big Data [4]. Sreenivas R. Sukumar reviewed the Machine Learning

Revised Manuscript Received on March 26, 2019

Mounica Vennapusa, Assistant Professor, Department of Information Technology, PACE Institute of Technology & Sciences.
Srikanth Bhyrapuneni, Assistant Professor, Department of Information Technology, PACE Institute of Technology & Sciences.

challenges and analyzed the challenges at scale to Machine Learning in Big Data era [5]. Omar Y. Al-Jarrah, described the Big Data challenges and introduced new algorithmic methods with less memory and processing to improve cost minimization to maintain its reliability and strength [6] Nagwa M. Elaraby, overviewed Deep Learning architectures and its applications, challenges on Big Data Analytics [7] Cheryl Ann Alexander, introduced techniques in machine learning; analyzed the technology improvement of Machine Learning applications for Big Data [8]. Alexandra L'Heureux, reviewed the issues that are highly relevant to Machine learning associated with Big Data and overviewed how emerging trends are responding to these issues [9]. Roheet Bhatnagar, presented a review of key challenges, latest developments in Machine Learning for Big Data Analytics [10], explored how Deep Learning context and how it can be used for addressing some Key challenges in Big Data Analytics [11].

III. SOME MACHINE LEARNING METHODS

Machine Learning algorithms can be divided into supervised, semi-supervised and unsupervised [18].

In Supervised learning, both the inputs and their desired outputs are provided and an algorithm is used to learn the mapping function from inputs to outputs[19]. Classification and regression are the two main tasks of supervised learning. In classification the output is to predict the target class while in regression the output is to predict the continuous values. K-NearestNeighbor (KNN), Support Vector Machine(SVM),DecisionTree(DT), Naive-Bayes(NB) are the algorithms for classification[36]. Polynomial regression and Linear regression are the algorithms for regression and in which algorithms like Neural Networks can be utilized for both classification and regression.

In unsupervised learning, only the input data is provided and the outputs are not known[36][38]. Clustering is the example of Unsupervised learning which includes the grouping of objects based on similarity criteria. K-means is the well-known example of this learning. Predictive analytics is based on Machine Learning to implement models used past data to predict the future. Algorithm such as neural networks, Naïve-bayes and Support Vector Machine can be used for this purpose.

Semi-supervised learning falls among Supervised and Unsupervised learning. In semi supervised learning both labeled and unlabeled data are used for training-generally a maximum of unlabeled data and a minimum measure of labeled data [18]. Typically semi supervised learning is opted only when the extracted labeled data requires skilled and related resources to learn from it.

Algorithm	Algorithm Type	Data Processing Tasks	Pros	Cons	Applications
Decision Tree [50]	Supervised learning	Classification or Regression	• Can handle both nominal and numerical attributes • Capable of dealing datasets that contain errors • Rapidly express complex alternatives clearly	• Sensitivity to the data set, to irrelevant attributes and to noise. • Require the target attribute to have only distinct values.	• Image Classification • Text Categorization
K-means [47]	Unsupervised learning	Clustering	• Computationally faster for big datasets, if k is small. • Produce tighter clusters, if the clusters are globular.	• Difficult to analyze K-Value, the number of clusters. • Sensitive to scale	• Document Classification • Customer Segmentation
Support Vector Machine [49]	Supervised learning	Classification or Regression	• Avoids overfitting • Easy to explicate results.	• For large datasets, training time is more. • Exhaustive for data of different types	• Pattern Recognition • Stock Market Forecasting
Neural Networks [52]	Semi-Supervised learning	Classification or Prediction	• Accurate predictive performance • Have tolerance to noisy data	• Determination of variable selection method. • Extrapolation	• Driving • Fraud Detectiont
Navie-Bayes Classifier [50]	Supervised learning	Classification or Regression	• Relatively simple and easy to use • Not sensitive to missing and noisy data	• Interprets all the data are equally important and independent. • Computational approximation is required.	• Spam Filtering • Sentimental Analysis
K-nearest neighbor algorithm [48]	Supervised learning	Classification or Regression	• Easy to understand and interpret • Gives more accuracy	• Computationally expensive • Usage of more memory	• Face Recognition • Medical Imaging Data

IV. MACHINE LEARNING IN BIGDATA

The following brief list identifies the different Machine Learning algorithms when applied in Big Data.

- **Decision tree** is a supervised learning technique which can be utilized either for classification or regression. In Decision Tree Learning, a training model is created and it identifies the outcome value by learning decision rules deduced from the data attributes [39]. These decision tree algorithms have multiple limitations in Bigdata. Firstly, whenever the data is very huge, creating a decision tree is very time taking. Secondly, the approach of data distribution is not maximized which leads to increase in the communication cost.
- **Support Vector Machine** is a supervised learning method which can be utilized either for classification or regression [40]. When applied to big data, the SVM technique is not efficient because of its high computational complexity. For huge amount of data, the requirement of computation and storage will gets increased massively.
- **K-Nearest Neighbor (KNN)** algorithms are used for regression and classification problem.KNN methods use a data and classify new data

points based on similar measures. The data is reserved to the class which has the most nearest neighbours. As the number of nearest neighbours gets increased, value of k, accuracy gets increased. For Bigdata applications, KNN is not realistic due to its high computation and memory usage cost.

- **Naive bayes** classifier is used strongly for classification task. It identifies membership probabilities for each class or data point that belongs to a certain class. The class with the highest probability is considered as the most likely class. In Bigdata because of the text redundant features and rough parameter estimation the performance of Naive Bayes is not feasible in text classification tasks.
- **Neural Networks**, a semi supervised technique used for both classification and regression. Neural Networks is a computing system consists of highly interconnected processing elements which processes data by their dynamic state response to external inputs [27]. One of the most popular algorithm of Neural Network is Back Propagation. For Bigdata with the increasing scale of information, neural networks have few challenges. Immense amount of data makes the technique difficult to maintain both reliability and effectiveness and also redundant data increases the workload of the system.

Algorithm	Accuracy	Tolerance to missing values	Tolerance to irrelevant data	Tolerance to noise	Handling overfitting	Speed
Decision Tree	Low	Medium	Medium	Low	Low	Fast
Support Vector Machine	High	Low	High	Low	Low	Fast
Neural Networks	Medium	Low	Low	Low	Low	Fast
Naive Bayes classifier	Medium	High	Low	Medium	Medium	Fast
K-Nearest Neighbour	Low	Low	Low	Low	Medium	Low

Table 2 Performance of Machine Learning algorithms on Bigdata

All these traditional machine algorithms possess different limitations for large amount of data. To overcome this, all these algorithms can be transformed into a series of Map and Reduce methods using Map Reduce programming model, such that these algorithms can produce time effectiveness and scalability [12-16].

MapReduce is an implementation for processing and generating large amounts of data. MapReduce programs are parallel in their nature, which supports for performing huge data analysis using number machines in the Cluster[46]. The basic MapReduce framework provides platforms like Hadoop for Machine learning. Hadoop is generally for storing, processing with large data set using Commodity hardware [4]. Hadoop consists of programming model known as MapReduce and a storage part known as Hadoop Distributed File System (HDFS)[41]. There are number of different vendors who pursue the way of Hadoop [44]. Apache Spark is the next following in the big data tools. Spark processes data much quicker than conventional disk processing [35]. This is an additional element for data analysts handling certain types of data to gain the faster outcomes .R programming tool is one of the widely used for statistical analysis of data among the open source big data tools. An R model constructed and tested on a local data source can be effortlessly implemented in different servers or even against a Hadoop information lake[53].

MapReduce does not aid for the repetition or particular key features needed to effectively iterate a MapReduce program. Software engineers constructing Machine Learning models on such frameworks need to develop looping in specialized ways exterior the core Map Reduce structure [43].

V. Advanced Machine Learning Methods for Big Data

For big data processing, most Machine Learning techniques are not inclusive. In other words, it is regularly need to use specific learning methods according to different data. When big data is concerned, it is in the need to scale up Machine Learning algorithms. The different learning methods are discussed below.

A. Deep Learning

Deep Learning also called Hierarchical Learning mainly uses Supervised and unsupervised Learning in deep architectures to learn hierarchical representations. The upside of Deep Learning is that the program assembles the data set by itself by unsupervision [7][21][27]. The two mainstream deep learning approaches are Deep Belief Networks (DBNs) and Convolutional Networks (CNNs) [45]. As the data is increasing, deep learning is giving predictive analytics solutions for large scale data sets with the increased computational and processing usage.

The recent advancements in regard to [34]: *Reviewed notable deep learning methods and that have been utilized for various Natural Language Processing tasks and provided its efficiency.*

B. Feature learning

High dimensional datasets have turned out to be increased which challenge the current learning to separate and categorise the meaningful data from the information. Feature learning a solution which can learn the useful representations of the data that makes simple and easy to adapt meaningful information. Feature learning or Representation learning is a collection of mechanisms that enables a computer program to naturally find the representations used for feature detection or classification from data collected and to learn the features and utilize them to access a specific problem [28].Feature selection, Feature extraction, and Distance metric learning are the three main types of Representation Learning.

The recent advancements in regard to [29]: *Proposed a novel approach to detect drive drowsiness using Feature Learning.*

C. Active Learning

There are circumstances in which unlabeled information is more such that manually labeling and getting labeled data is expensive [22]. In such situations, learning algorithms can question the user for labeled data. This type of iterative supervised learning is known as Active learning. Active learning is a learning method which is able to interact and ask the user to get the impulse outcomes at new data points. Since the user chooses the samples, the number of samples to learn a concept can be much less than the number required in normal supervised learning.

The recent advancements in regard to [30] : *Proposed methods based on active learning for efficient hybrid Biophysical variable retrieval.*

D. Ensemble Learning

The amount of information created by business, social networks and other domains has increased massively. All these data are useful only if it is accurately performed so that users can produce appropriate resolution based on them. Rather than making one model and desiring that model as the best and realistic predictor we can make, ensemble methods take a group of models into account, and will aggregate those models to give a final model. Ensemble Learning to configure a collective whole where several methods are more efficient than an isolated learning method [24][26]. Bagging, boosting, stacking, and error-correcting output are the four types of Ensemble Learning.

The recent advancements in regard to [23]: *Proposed an SVM algorithm, for breast cancer diagnosis to increase diagnosis accuracy and to reduce the variance based on Ensemble Learning.*

E. Transfer Learning

In current Machine Learning scenarios, the collection of trained data is expensive or difficult. This has given the requirement to generate more productive learners trained with more easily accessible data from distinct domains. This learning is known as Transfer learning. Transfer learning is the capability of a system to accept and apply knowledge and experience learned in past to present tasks [25].If the data from different feature spaces are to be processed and have different distributions, transfer learning will be the optimum solution to solve new problems. Inductive transfer learning, Transductive transfer learning, and Unsupervised transfer learning are the different categories of Transfer Learning.

The recent advancements in regard to [31]: *Introduced an automatic transfer learning (ATL), a transfer learning framework for Bigdata.*

F. Online Learning

Data is being produced in immense amounts all over the place. Batch learning algorithms take batches of training data to train a model. In contradiction an On-line learning algorithm, speculate model and afterward grab one-one perception from the preparation and recalibrates the weights on each information parameter. Besides, as it doesn't require all information to be available, this technique provides an alternative for data availability and locality. If the dealing with real-time data is needed, then Online learning will be the suitable solution.

The recent advancements in regard to [32]: *Introduced an efficient methodology of online learning machines to achieve a result for Big-Data streaming applications.*

G. Continuous Learning

Continuous Learning also referred as Lifelong Learning. It is an adaptive Machine Learning approach that learns continuously, gathers the knowledge learned in the past, and utilizes to help for future learning and problem solving. Lifelong learning is implemented similar to human learning. In Continuous learning, learning is continuous; knowledge is retained and used to solve different problems. Lifelong learning labels situations when a user faces a stream of learning tasks.

The recent advancements in regard to [33]*: Surveyed knowledge-based topic methods and lifelong learning methods for Natural Language processing (NLP) with huge data sets.*

H. Distributed Learning

There is frequently exciting information covered up in massive volumes of data. Learning from these new data has the incompetence of learning techniques to utilize all the data to learn within a specific amount of time. In this position, distributed learning aims to be an optimistic solution, whereas allotting the learning process between various applications is a characteristic method for scaling up learning Algorithms. For huge amount of data, distributed and parallel learning methods have more grounded focal points.

The recent advancements in regard to [20]*: Proposed MetaLP, a distributed statistical modeling framework where statistical modeling meets big data processing which is suitable for large-scale data analysis.*

VI. ISSUES AND CHALLENGES

- Data with more heterogeneous, under-represented classes and impaired with noise, may be hard to process by an automatic learning algorithm. Although the previous mentioned techniques such as transfer learning with high-performance learners can easily obtain data from different domains, are greedy and slow. Due to the extensive size of the data attribute, it is additionally a complex and non productive one to pre-process it.
- One of the incredible difficulties in a wide assortment of learning issues is the capability to acquire adequate labeled data for modeling purposes. Labeled information is regularly costly to acquire, and requires more human effort. Although Active learning obtained the desired outputs at new data points but with this method, there is a hazard that the algorithm is overpowered by uninformative models.
- In terms of various data tasks, types, and features, the required learning mechanisms are distinct, even a Machine Learning

technique support is used for big data processing. Accordingly, to upgrade the efficiency of data processing, the union of Machine Learning with another technique has to be proposed.

- If the data size is huge, it is difficult to process the data either by online or distributed learning. Online learning in big data needs more amount of time for training on an isolated system. Distributed learning with a more number of systems minimizes the retained effectiveness per system and affects the whole performance.
- Big data often obtains wide range of inputs, high dimensionality data sets, and more varieties of outcomes. These attributes lead to high complexity of running time and proposed models. Deep learning mechanisms with a central processor and storage face a challenge in dealing with these properties.

VII. CONCLUSION

This paper has provided a systematic study of the traditional Machine Learning mechanisms and their comparative study when applied with Bigdata. As these algorithms are not inclusive, advanced learning methods are discussed with their issues and challenges. Various techniques have been implemented to process Machine Learning algorithms to access large scale data such as MapReduce and distributed frameworks such as Hadoop. Advanced methods include several mechanisms in which Deep learning has the potential to conquer the difficulties of Machine Learning with Big Data. Deep learning has the capability in dealing and learning problems found in huge volumes of input data in spite of having few challenges. The progressive learning and extraction of various levels of data abstractions in Deep Learning gives a certain degree of explication for Big Data Analytics.

Highlight a section that you want to designate with a certain style, and then select the appropriate name on the style menu. The style will adjust your fonts and line spacing. Do not change the font sizes or line spacing to squeeze more text into a limited number of pages. Use italics for emphasis; do not underline.

REFERENCES

1. Lidong Wang, "Data Mining, Machine Learning and Big Data Analytics" , International Transaction of Electrical and Computer Engineers System, 2017, Vol. 4, No.2, 55-61
2. Ming Ke, Yuxin Shi, "Big Data, Big Change: In the Financial Management", Open Journal of Accounting, 2014, 3, 77-82
3. Yuhua Xu and Shuo Feng ,Junfei Qiu, Qihui Wu, Guoru Ding, "A survey of machine learning for big data Processing", EURASIP Journal on Advances in Signal Processing (2016) 2016:67
4. Junfei Qiu and Youming Sun, "A Research on Machine Learning Methods for Big Data Processing", International Conference on Information Technology and Management Innovation (ICITMI 2015)
5. Sreenivas R. Sukumar, "A Research on Machine Learning Methods for Big Data Processing", Conference Paper August 2014.
6. Omar Y. Al-Jarrah , Paul D. Yoo,Sami Muhaidat,George K. Karagiannidis,Kamal Taha, "Efficient Machine Learning for Big Data: A Review", Big Data Research2(2015) 87–93
7. Nagwa M. Elaraby, Mohammed Elmogy, Shereif Barakat, "Deep Learning: Effective Tool for Big Data Analytics", International Journal of Computer Science Engineering (IJCSE) ISSN : 2319-7323 Vol. 5 No.05 Sep 2016
8. Lidong Wang, Cheryl Ann Alexander, "Machine Learning in Big Data", International Journal of Mathematical, Engineering and Management Sciences Vol. 1, No. 2, 52–61, 2016
9. Alexandra L'Heureux, Katarina Grolinger, Hany F. ElYamany, Miriam A. M. Capretz, "Machine Learning with Big Data: Challenges and Approaches", IEEE Access April 2017
10. Roheet Bhatnagar, "Machine Learning and Big Data Processing: A Technological Perspective and Review", Springer International Publishing AG, part of Springer Nature 2018
11. Maryam M Najafabadi,,Flavio Villanustre,Taghi M Khoshgoftaar,Randall Wald, Edin Muharemagi, "Deep learning applications and challenges in big data analytics", Najafabadi et al. Journal of Big Data (2015)
12. Pwint Phyu Khine, Wang Zhao Shun1, "Big Data for Organizations: A Review", Journal of Computer and Communications, 2017, 5, 40-48
13. Anushree Priyadarshini and SonaliAgarwal, "A Map Reduce based Support Vector Machine for Big Data Classification", International Journal of Database Theory and Application Vol.8, No.5 (2015), pp.77-98
14. Tamer Tulgar, Ali Haydar and Ibrahim Ersan, "A Distributed K-Nearest Neighbor Classifier for Big Data", BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 6, No. 2, April 2018
15. Wei Dai, Wei Ji, "A Map Reduce approach of c4.5 Decision tree Algorithm", International Journal of Theory and Application; vol 7 no.1(2014), pp 49-60
16. Kairan Sun, Xu Wei, Gengtao Jia, Risheng Wang, and Ruizhi Li, "Large-scale Artificial Neural Network:MapReduce-based Deep Learning", arXiv:1510.02709v1 [cs.DC] 9 Oct 2015
17. .Available [online] https://www.forbes.com/sites/bernardmarr/2018
18. PETER HARRINGTON, "Machine Learning in Action"
19. Annina Simon, Mahima Singh Deo, Mahima Singh Deo, S. Venkatesan, D.R. Ramesh Babu,D.R. Ramesh Babu, "An Overview of Machine Learning and its Applications", International Journal of Electrical Sciences & Engineering (IJESE);Vol1, Issue 1; 2015 pp. 22-24
20. Scott Bruce, Zeda Li, Hsiang-Chieh Yang and SubhadeepMukhopadhyay, "Nonparametric Distributed Learning Architecture for Big Data: Algorithm and Applications", rXiv:1508.03747v5 [stat.AP] 26 Feb 2018
21. Mehdi Gheisari, Guojun Wang, "A Survey on Deep Learning in Big Data", 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference
22. Charu C. Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, "Active Learning: A Survey"
23. Haifeng Wang, Bichen Zheng, , Sang Won Yoon, Hoo Sang Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis", European Journal of Operational Research 267 (2018) 687–699
24. Heitor Murilo Gomes,Jean Paul Barddal, Fabr,Icio Embreck,Albert Bifet, "A Survey on Ensemble Learning for Data Stream Classification", Article in ACM Computing Surveys · March 2017
25. Ricardo Sousa, Luis M. Silva, Luis A. Alexandre, Jorge Santos, "Transfer Learning: Current Status, Trends and Challenges"
26. R. Anandan, Srikanth Bhyrapuneni, K. Kalaivani, P. Swaminathan, "A survey on big data analytics with deep learning in text using machine learning mechanisms", International Journal of Engineering & Technology, 7 (2.21) (2018) 335-338
27. Pariwat Ongsulee, "Artificial Intelligence, Machine Learning and Deep Learning", 2017 Fifteenth International Conference on ICT and Knowledge Engineering
28. Zhong G, Wang LN, Ling X, Dong J, "An Overview on Data Representation Learning: From Traditional Feature Learning to Recent Deep Learning", The Journal of Finance and Data Science (2017), doi: 10.1016/j.jfds.2017.05.001
29. Kartik Dwivedi, Kumar Biswaranjan and Amit Sethi, "Drowsy Driver Detection using Representation Learning"
30. Jochem Verrelst, Sara Dethier, Juan Pablo Rivera, Jordi Muñoz-Marí, Gustau Camps-Valls, and José Moreno, "Active Learning Methods for Efficient Hybrid Biophysical Variable Retrieval; IEEE Geoscience And Remote Sensing Letters".
31. Karl Weiss , Taghi M. Khoshgoftaar and DingDing Wang, "A survey of transfer learning", Weiss et al. J Big Data (2016) 3:9
32. Karim Kanoun, Cem Tekin, David Atienza, and Mihaela van der Schaar, "Big-Data Streaming Applications Scheduling Based on Staged Multi-armed Bandits",

IEEE Transactions On Computers, Vol. Xx, No. Xx, Month 20xx
33. M. Taimoor Khan, Mehr Durrani, Shehzad Khalid and Furqan Aziz, "Lifelong aspect extraction from big data: knowledge engineering", Khan et al. Complex Adapt Syst Model (2016) 4:5
34. Tom Young, Devamanyu Hazarika, Soujanya Poria , Erik Cambria, "Recent Trends in Deep Learning Based Natural Language Processing", arXiv:1708.02709v7 [cs.CL] 10 Oct 2018
35. D. Singh and C. K. Reddy, "A Survey on Platforms for Big Data Analytics," Journal of big data, vol. 2, no. 1, pp. 1–20, 2015
36. TM Mitchell, "The discipline of machine learning", (Carnegie Mellon University,School of Computer Science, Machine Learning Department, 2006)
37. TH Davenport, P Barth, R Bean, "How "big data" is different", MIT Sloan Manage Rev 54(1), 22–24 (2012)
38. TM Mitchell, Machine learning (McGraw-Hill, New York, 1997)
39. Yu-Wei CD. "Machine learning with R cookbook", Packt Publishing Ltd; 2015, Mar 26
40. Wikibook, "Data Mining Algorithms In R - Wikibooks, open books for an open world", PDF generated using the open source mwlib toolkit. See http://code.pediapress.com/, 2014 14 Jul
41. T. White, "Hadoop: the Definitive Guide", O'Reilly Media Inc., California, 2009
42. M. Rouse, "Machine Learning Definition," 2011. http://whatis.techtarget.com/definition/machine-learning
43. K. Grolinger, M. Hayes, W. A. Higashino, A. L'Heureux, D. S. Allison, and M. A. M. Capretz, "Challenges for MapReduce in Big Data," in Proceedings of the 2014 IEEE World Congress on Services (SERVICES), 2014, pp. 182–189
44. https://towardsdatascience.com/8-open-source-big-data-tools-to-use
45. Xue-wen Chen and Xiaotong Lin, "Big Data Deep Learning: Challenges and Perspectives," IEEE Access, vol. 2, pp. 514–525, 2014
46. Josep Lluis Berral-García, "A Quick View on Current Techniques and Machine Learning Algorithms for Big Data Analytics", 978-1-5090-1467-5/16/$31.00 ©2016 IEEE
47. Andreopoulos B, "Literature Survey of Clustering Algorithms",Workshop, Department of Computer Science and Engineering,York University, Toronto, Canada, 2006 June 27
48. Clark M, "An introduction to machine learning: with applications in R", University of Notre Dame, USA, 2013.
49. Andreopoulos B, "Literature Survey of Clustering Algorithms", Workshop, Department of Computer Science and Engineering, York University, Toronto, Canada, 2006 June 27
50. Kotsiantis, S. B. (2007), "Supervised machine learning: a review of classification techniques",Informatica,31, 249-268
51. A short history of machine learning - every manager shouldread. https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learningevery-manager-should-read/2/#28d56abd6b1b.Accessed 10 June 2017. Dong, X., Li, Y., Wu, C., Cai, Y, "A learner based on neural network for cognitive radio", In: 2010 12th IEEE International Conference on Communication Technology(ICCT), pp. 893–896. IEEE (2010)
52. Madala, S. R., Rajavarman, V. N., & Vivek, T. V. S. (2018). Analysis of Different Pattern Evaluation procedures for Big Data Visualization in Data Analysis. In Data Engineering and Intelligent Computing (pp. 453-461). Springer, Singapore.

AUTHORS PROFILE

Author-1 Photo

Mounica Vennapusa, Assistant Professor, Department of Information Technology, PACE Institute of Technology & Sciences.

Author-2 Photo

Srikanth Bhyrapuneni, Assistant Professor, Department of Information Technology, PACE Institute of Technology & Sciences.