ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter138892020.pdf
https://doi.org/10.30534/ijeter/2020/138892020

Trending Algorithms in Machine Learning and issues along with Big data Context in real time Data Processing

S. Nagendram[1], M.S.R. Kiran Nag[2], S.S.S. Kalyan[3], Narendra .S[4]
[1]Associate Professor, Department of Electronics and Communication Engineering, K L E F, Guntur, AP, India.
[2]Research Scholar, Department of Computer Science and Engineering, Andhra University,Vizag, AP, India.
[3]Assistant Professor, Department of Electronics and Communication Engineering, K L E F, Guntur, AP, India.
[4]Assistant Professor, Department of Computer Science and Engineering, UCE, Guntur, AP, India.
reena1286@gmail.com[1], kirannag02@gmail.com[2], ssskalyan@kluniversity.in[3], narendra.smile@gmail.com[4]

ABSTRACT

Machine Learning (ML) is booming up and mandatory for many domains like banking, retail social media and online shopping platforms like amazon and flip kart. The article describes the importance of various algorithms along with the fundamental concept with simple code and the corresponding issues. The other dimension of this work is summary of various applications based on the algorithms and some of the observations which may help the researchers in the area of ML to come up with suitable methodologies and corrective measures. The work involves the description of regression, classification and recommender systems with suitable snippets of the coding in R programming. The discussion also provides various issues identified in the regression, classification and recommender systems which may be leading to the future scope for the researchers in the area of ML. We believe that the work helps to know the basic algorithms in the said context with the application areas and issues based on the application domains like social media (Facebook, Twitter), banking sector classification of the customers based on the Cibil score and recommender systems along with product recommendation to the customers while working with amazon and flip kart kind of the online shopping sites. The outcome of the work is a clear specification of algorithms with the help of R programming and working with issues which leads to the new research trends in ML. Integration of Hadoop and R is the best combination in the field of distributed computing and analytics point of view. The algorithms regression and classification along with the specific packages allows making use of the algorithms in the simple manner.

Key words: Machine Learning, Regression, R, Classification, Recommender Systems.

1. INTRODUCTION

Huge data processing is always a critical issue in the social media, online shopping and other applications such as banking, retail etc. The huge amounts of the data storage are tedious and processing the same is cumbersome. The IT companies need to answer two important questions like whether the existing storage servers are ready to handle the current data storage demands and the current algorithms are enough to process such huge data[1], [2]. Whatever the answer by the companies but the ultimate solution is Hadoop Distributed File System (HDFS) along with Map Reduce (MR). The HDFS is suitable to handle huge data by following block based storage and MR is a parallel and distributed environment where the processing of the data is undergone into various steps like Mapper, reducer and Driver aspects. ML anyhow is used to improve the performance of the task in a time bound related to any application [3]. ML provides various algorithms like Regression, Classification and recommender systems so as to uncover the required data and allows the developers to analyze the data in a versatile manner. The flow of the work is in Section II the big data scenarios and issues were described, In Section III the ML algorithms and corresponding simple code along with results were described, In Section IV integration of big data and ML were described along with available architectures of RHDFS and RMR. In Section V the conclusion and future scope were described.

2. BIGDATA VENDORS AND OTHER ISSUES IN THE EXISTING ARCHITECTURES

The term big data is used everywhere now, is basically dealing with huge amounts of the data, and

data itself becomes one problem in the handling of the application [4]-[6]. In the introduction mentioned earlier two importance factors that guide the big data scenarios are storage and processing of the data. Big data is a kind of the problem but the solution to that big data problem is available in various formats, some of the vendors here worth to mention and they are all trying to contribute to the big data problems by providing solutions[7]. Hadoop is a framework which provides storage and processing logic for big data problems. Figure 1 shows the contribution of the vendors in the market of Hadoop with the potential percentage of performance in the current market.

Figure 1: Various vendors of Hadoop (source: dezyre lectures)

Apache Hadoop , Cloudera, AWS, MapR, Hortonworks and IBM are leaders in the market in the context of integration of various tools like Pig Latin, Hive, HBase and support of ML

Libraries in the eco system [8]. Issues in the architecture involves the Hadoop administration kind of the aspects like there is no common procedure to install and work with Hadoop frame work, for example apache Hadoop provides all the .tar files and manually the user need to configure the files like .bashrc, core-site.xml, apred-site.xml,Hadoop-env.sh. In case of the Cloud era a complete GUI based architecture is available but the problem is the researchers or developers are not having any idea about the internal files and other daemons which are playing a vital role inherently [9]. So the solution is the process must be same and fewer complexes and the configuration can be customized according to the specific need of the installation. There is myth like Hadoop installation and learning is difficult which is not true, Hadoop framework is easy to learn and apply on the huge data sets [10].

3. ML REGRESSION AND CLASSIFICATION WITH R IMPLEMENTATION

Machine Learning is a process of improving performance of a task with in a time specification, with ML the process can be improved every time after running the task with the greater performance. ML is a data-driven approach various algorithms are available but the current wok dealing with regression, classification and recommender systems only[11]-[13].

➢ x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
➢ y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
➢ relation <- lm(y~x)
➢ png(file = "linearregression.png")
➢ dev.off()
➢plot(y,x,col = "blue",main = " Linear Regression",abline(lm(x~y)),cex = 1.3,pch = 16,xlab = "Weight",ylab = "Height")

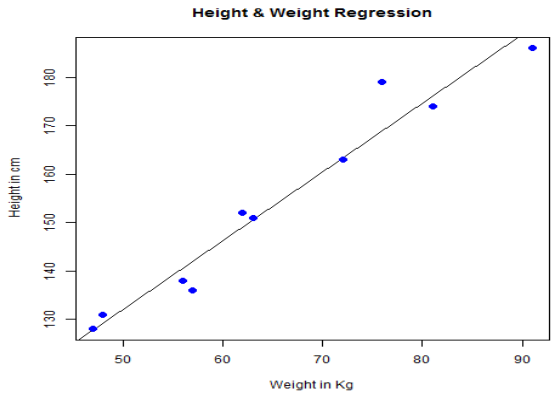

Figure 2: Regression model output

Regression draws a line and establishes a kind of correlation between the identified values [14]. Regression in the code is simple and gives a common relation but there is a possibility of the revision in the model so as to explore the model in detailed manner. Classification is most commonly used method in the ML models the code snippet implements the usage of classification with rpart package. Classification model involves decision trees, random forest [15].

➢ setwd("C:\\Users\\haarshadatta\\Desktop\\uma")
➢ getwd()

```
➢ data1<-
read.csv("Mail_Respond.csv",head=TRUE)
➢ library(rpart)
➢ library(rpart.plot)
➢ mymodel<-
rpart(outc~dist+house+inc+pc,method='class',control
= rpart.control(minsplit = 10))
➢ mymodel
➢ plotcp(mymodel)
➢ prop.table(myt)*100
```

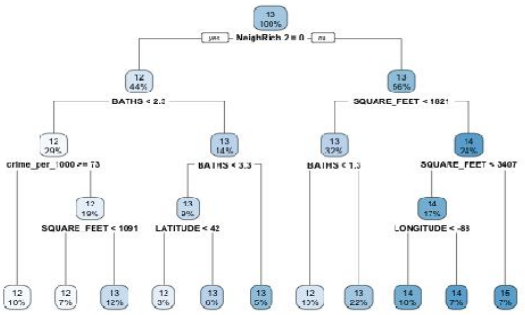

Figure 3: Classification Model of Housing data.

The issue in case of classification model is basic classification is distinct with decision tree and random forest where there is a scope of integrated classifier model which provides a parameter based approach for selecting the specific model at the time of implementation[16]. Recommender systems deals with profile based and content based outcome [17]. The best example is in case of the amazon recommender system at the moment we select the product like laptop then the application also recommends the products within the same cost and configuration kind of the settings, the process of identifying the products and recommendation of the products comes under the category of ML algorithms [18].

4. ML AND HADOOP INTEGRATION WITH AVAILABLE TOOLS

Huge data storage is the benefit of Hadoop, applicability of the various algorithms and getting the valuable analytics is the advantage of ML(R provides various packages like rpart, e1071, kmeans…)[19].With R programming the implementation of ML algorithms like classification, clustering and recommender systems, twitter analytics and sentiment analysis along with social media analytics are possible. The problem with R is while performing ML algorithms it has to dependent on RAM only which obviously limits the performance of the algorithm [20]. So as to solve the memory problem R has to depend on Hadoop. Hadoop is not meant for analytics but for storage and processing so the integration of R and Hadoop resolves the issue of storage and analytics requirements [21], [22]. RHadoop provides various packages to manage and analyse data with Hadoop. rhdfs is a package where through R we can perform reading, writing and modify the files stored in HDFS with basic connectivity. Rhbase is used to establish a NoSQL environment without indexing, without normalization of the data [23]-[25]. HBase provides a quorum based architecture to create, insert and read the data. Plymr is a package where user can perform some manipulations on the data. Rmr2 is a package which allows performing statistical analysis in R via Map Reduce [26]-[28].

5. CONCLUSION AND FUTURE SCOPE

The work aims in the Hadoop and ML integration. The description of the Hadoop and the corresponding issues in case of the unified process of configuring the cluster has been outlined. The ML algorithms like regression and classification with R implementation has been outlined and the corresponding results have been displayed. The integration of the Hadoop and R is vital requirement which solves the storage issue of the R and analytics issue of the Hadoop. The future scope of the work is a framework to install Hadoop with common steps related all the vendors mentioned in the Section I of the work. The other future scope is simple integration of R and Hive, R and Pig as well as R and Flume, R and Sqoop which greatly reduces the complexity of the development.

REFERENCES

1.Srinivasa Rao Y., Ravikumar G., Kesava Rao G., Syed M.S. (2017),**'Interconnected transmission line fault detection using wavelet transform and a novel machine learning algorithm'**,*Journal of Advanced Research in Dynamical and Control Systems*,9(12),PP.142-150.

2. Ayushree, Arora S.K. (2017),**'Comparative analysis of AODV and DSDV using machine**

learning approach in MANET',*Journal of Engineering Science and Technology*,12(12),PP.3315-3328.

3. Sripath Roy K., Roopkanth K., Uday Teja V., Bhavana V., Priyanka J. (2018) , **'Student career prediction using advanced machine learning techniques'**,*International Journal of Engineering and Technology(UAE)*, 7 (2),PP. 26- 29

4. Uma Pavan Kumar K, **Various Issues in Hadoop Distributed File System, Map Reduce and Future ResearchDirections,** *International Journal of Pure and Applied Mathematics*, Volume 120 No. 6 2018, 4441-4451, June 24, 2018.

5. *www.dezyre.com*

6. Rama Rao K.V.S.N., Sivakannan S., Prasad M.A., Agilesh Saravanan R. (2018) , **'Technical challenges and perspectives in batch and stream big data machine learning'**,*International Journal of Engineering and Technology(UAE)*, 7 (1),PP. 48- 51.

7. Harish Balaji, Ujjwal Pal and UmaPavan Kumar K., **Big data Techniques and Analytics in Distributed E-commerce business,** *International Journal of Control theory and applications*, Volume: No.9 (2016) Issue No.:3 (2016), Pages : 1719- 1726.

8. Ayushree, Balaji G.N. (2018) , **'Comparative analysis of coherent routing using machine learning approach in MANET'**,*Smart Innovation, Systems and Technologies*, 77 (),PP. 731- 741

9. S. Madden**, "From Databases to Big Data.,"** *IEEE Internet Comput.*, vol. 16, no. 3, 2012.

10. P.Zikopoulos, C. Eaton, and others, **Understanding big data: Analytics for enterprise class hadoop and streaming data.** *McGraw-Hill Osborne Media*, 2011.

11. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, **"Big data,"** *Manag. Revolut. Harv. BusRev*, vol. 90, no. 10, pp. 61–67, 2012.

12. R. Appuswamy, C. Gkantsidis, D. Narayanan, O. Hodson, and A. Rowstron, **"Scale-up vs Scale-out for Hadoop:Time to rethink?,"** *in Proceedings of the 4th annual Symposium on Cloud Computing*, 2013, p. 20.

13. Danthala S., Rao S., Mannepalli K., Shilpa D. (2018) , **'Robotic manipulator control by using machine learning algorithms: A review'**,*International Journal of Mechanical and Production Engineering Research and Development*, 8 (5),PP. 305- 310

14. C. P. Chen and C.-Y. Zhang, **"Dataintensive applications, challenges, techniques and technologies: A survey on Big Data,"** *Inf. Sci.*, vol. 275, pp. 314– 347, 2014.

15. T. B. Murdoch and A. S. Detsky, **"The inevitable application of big data to health care,"** *Jama*, vol. 309, no. 13,pp. 1351– 1352, 2013.

16. I. Mashal, O. Alsaryrah, and T.-Y. Chung, **"Performance evaluation of recommendation algorithms on Internet of Things services,"** *Phys. Stat. Mech. Its Appl.*, vol. 451, pp. 646–656, 2016.

17. K. Shvachko, H. Kuang, S. Radia, and R. Chansler**, "The hadoop distributed file system,"** *in 2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, 2010, pp. 1–10.

18. J. Dean and S. Ghemawat, **"MapReduce: simplified data processing on large clusters," Commun.** *ACM*, vol.51, no. 1, pp. 107–113, 2008.

19. J. Y. Monteith, J. D. McGregor, and J. E. Ingram, **"Hadoop and its Evolving Ecosystem.,"** in *IWSECO@ICSOB*, 2013, pp. 57–68.

20. Raghav, R. S.; Amudhavel, J.; Dhavachelvan, P. "**A Survey on Tools used in Big Data platform".** *Advances and Applications in Mathematical Sciences.* NOV 2017.

21. S. Hoffman, **Apache Flume: Distributed Log Collection for Hadoop.** *Packt Publishing* Ltd, 2013.

22. Ramanujam S.S., Sivaneshwar P., Naren J., Madhumitha S., Vithya G. (2019**), 'A study on hybrid recommender system with deep learning and deployment in big data'**, *Test Engineering and Management*, 81(44147), PP.1869-1875.

23. Sai, M. Krishna; Sivaramakrishna, N.; Teja, P. V. N. S. Ravi; Prakash, Kolla Bhanu "**A Hybrid Approach for Enhancing Security in IOT using RSA Algorithm"** *HELIX 2019.* 10.29042/2019-4758-4762.

24. C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, **"Pig latin: a notso-foreign language for data processing,"** *in Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp.1099–1110.

25. Bashir, Ali Kashif; Arul, Rajakumar; Basheer, Shakila; Raja, Gunasekaran; Jayaraman, Ramkumar; Qureshi, Nawab Muhammad Faseeh." **An optimal multitier resource allocation of cloud RAN in 5G using machine learning"** *Transactions on Emerging Telecommunications Technologies.* Aug 2019. *10.1002/ett.3627.*

26. S., Sai Anil P., Pavan E.V.S., Amarendra V. (2019), **'Performance evaluation of wide area network using cisco packet tracer',** *International Journal of Advanced Trends in Computer Science and Engineering*, 8(6), PP.2915-2919.

27. Arba Asha Altaye and Dr. J. Sebastian Nixon**" A Comparative Study on Big Data Applications in Higher Education"** *International Journal of Emerging Trends in Engineering Research (IJETER)* Volume 7 No. 12 (2019).

28.Edward Chandra, Pangondian Prederikus, Stefanie Liu and Gunawan Wang." Cukur361 Mobile Application Design for SME using Hadoop Framework", *International Journal of Emerging Trends in Engineering Research (IJETER)* Volume 7 No. 12 (2019) .