Big Data Predictive Analysis:Using R Analytical Tool

Priyanka P. Shinde
Assistant Professor, Department of MCA
Government College of Engineering Karad,
Karad, Maharashtra, India
priyanka.shinde@gcekarad.ac.in

Kavita S. Oza
Ass. Professor, Computer Science Department
Shivaji University, Kolhapur
Kolhapur, Maharashtra, India
skavita.oza@gmail.com

R. K. Kamat
Professor, Department Of Electronics
Shivaji University, Kolhapur
Kolhapur, Maharashtra, India
rkk_eln@unishivaji.ac.in

***Abstract* - Data available in large volume, variety is generally termed as Big Data. Since Big data is difficult to analyze using traditional data processing techniques, many new data processing tools and techniques have evolved over the need to practice result-oriented big data analysis. In this paper, big data has been analyzed using one of the advance and effective data processing tool known as R Studio to depict predictive model based on results of big data analysis. Couples of algorithms - Random Forest (RF) and Latent Dirichlet Allocation (LDA) are applied over R package in order to find out more concrete results. To portray operational demonstration of this model, author has performed case study by analyzing fertility associated big data and come up with predictive model which will help to foretell certain possibilities well in advance.**

Keywords—Big Data, R, Random forest, Latent Dirichlet Allocation, Predictive analytics, Healthcare

I. INTRODUCTION

Data is collection of facts, stats and information together which can help one to understand particular area in depth. As the term 'Big Data' itself suggests that its collection of large volume of data together, it has characteristics like Volume, Variety, and Veracity [1,2] etc. and often available in unstructured format. Hence, available big data need to be systematized to find out interesting facts and figures and work out on this data with specific strategies which can result in some concrete outcome. Many new tools and techniques have discovered to operate over big data which helps to perform big data analysis effectively. In conjunction with tools and techniques, one can apply different kinds of algorithm for betterment of big data analysis.

There are many different fields like social media, education, agriculture, healthcare, science etc. which are generating large volume of data that known as big data. In the Healthcare area there were many researches conducted but this is the field which need to be work more and more since there are drastic changes in human lifestyle, environment, food etc. So it's need of the time that have to find better treatment plans, better diagnosis, less cost for medicines, health plan etc.[3,4]. This paper presents experiment conducted using R tool on healthcare data. Section-I gives brief information related to R Analytical tool, Section-II describes implementation of algorithm, Section-III presents case study having fertility related data, source code in R, Section IV gives results , Section V describes conclusion and future scope.

II. R ANALYTICAL TOOL

R Open source software, statistical tool, provides facility of not only data analytics but also visual analytics, strong graphical support, and now towards predictive analytics to get better decision making and solution finding [5] .

R is also a programming language having CRAN which provides different packages and functions, user can develop own functions and packages. Program can link with C, C++, FROTRAN etc. and call at execution time[6]. It is interpreted language. Some features of R are

1. Open Source Software
2. Graphical and statistical tool
3. Simple and effective programming language.
4. Availability of large data handling, manipulation and storing facility
5. Visual analytics
6. Supports other programming language.

Many industries have used R tool to make effective decision with progressive business planning. Some of them are as follows[7].

TABLE I. R USED IN INDUSTREIS

Company Name	Use of R
Google	To find out ROI of advertising
Face book	Analysis of face book status updates
Microsoft	Xbox matchmaching service, statistical with Azure ML framework
Ford Motor	For statistical analysis and data driven decision support
John Deere	Time series modeling and geospatial analysis
Lloyds	Develop motion chart to provide analysis to investors

III. IMPLEMENTED ALGORITHEMS

Following are the algorithms which are implemented to get better outcomes.

A. Random Forest

The heart of classification is Random forest algorithm which consists of number of decision trees. The concept "Bagging" come forward for feature selection to assemble a decision tree with variance. All trees are developed as [8]:

1. If the number of cases for training set X, sample randomly X cases having replacement from main data. This sample used for the training set for growing the new tree
2. Let their are Y number of input variables, variable 'n' select as it is smaller than Y(i.e. n<Y) at each node. From Y select variable n randomly, and select best split in n variable. The value of 'n' is held constant during the forest growing.
3. Each tree is grown to largest extent possible without pruning.

Random forest can be implemented efficiently on huge data sets. It is the leading algorithm used in classification for better accuracy[9].

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an algorithm used for statistical model that allows number of observations to be explained by unobserved groups which explains causes of similarity of data. LDA is an algorithm which helps to analyze different kind of data like images, audio, text files etc. LDA aims to find short imagery for variables in a data collection. LDA helps to find out facts and figures with step by step exploration of data. LDA suggests that one way of concise the content of a document quickly is to look at the set of style it uses[10]. Graphically it will represent as follows, boxes i.e. plates are nothing but replicates. Document is shown in outer plate, repeated choice of topics, words within document shows in inner plate [11].

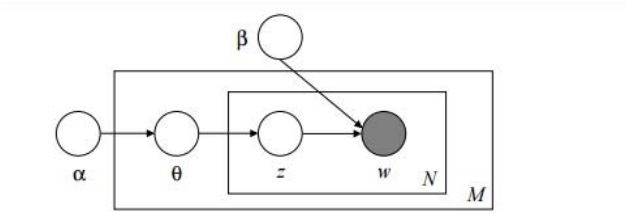

FIGURE 1. GRAPHICAL MODEL REPRESENTATION OF LDA FIGUR FROM[11]

IV. CASE STUDY

In proposed research author have used health related data having information related to fertility[12]. Different attributes and possible values are mentioned below.

TABLE II. DATASET DESCRIPTION

Attributes	Possible Values	Values in data set
Season in which the analysis was performed	1) winter, 2) spring, 3) Summer, 4) fall	(-1, -0.33, 0.33, 1)
Age at the time of analysis	18-36	(0, 1)
Childish diseases ie , chicken pox, measles, mumps, polio	1) yes, 2) no.	(0, 1)
Accident or serious trauma	1) yes, 2) no.	(0, 1)
Surgical intervention	1) yes, 2) no	(0, 1)
High fevers in the last year	1) less than three months ago, 2) more than three months ago, 3) no.	(-1, 0, 1)
Frequency of alcohol consumption	1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never	(0, 1)
Smoking habit	1) never, 2) occasional 3) daily	(-1, 0, 1)
Number of hours spent sitting per day	ene-16	(0, 1)
Diagnosis	normal (N), altered (O)	(N,O)

Data is available in CSV (comma separated value) file format. In R Studio read the data in .csv file, then clean data by removing NA (not applicable) fields, removing unwanted columns from data set. Caret(Classification and Regression

Training) package: main use to create predictive model. By using this package author have conducted splitting of data as training data and testing data.

R Source Code :

```
trainingnew <- read.csv("~/R/dataset.csv", header=FALSE)
View(dataset)
rmdsTrain<-dsTrain[, apply(dsTrain, 2, function(x)  !any(is.na(x)))]
cleandsrain<-rmdsTrain[,-c(5:6)]
dim(cleandsTrain)
library("caret", lib.loc="~/R/R-3.3.1/library")
inTrain <-createDataPartition(y=cleandsTrain$V10, p=0.60,
list=FALSE)
trainingData<-cleandsTrain[inTrain,]
testingData<-cleandsTrain[-inTrain,]
library(caret)
set.seed(1234)
cleandstest<-dsTest[,names(cleandsTrain[,-8])]
dim(cleandstest)
```

By implementing code data is uploaded in R shown in following figures.

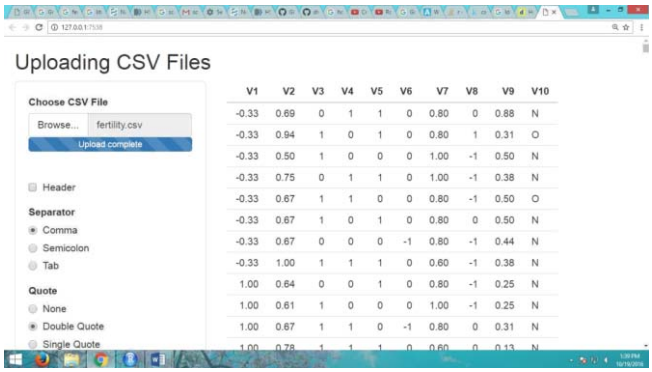

FIGURE 2. ORIGINLAL DATASET

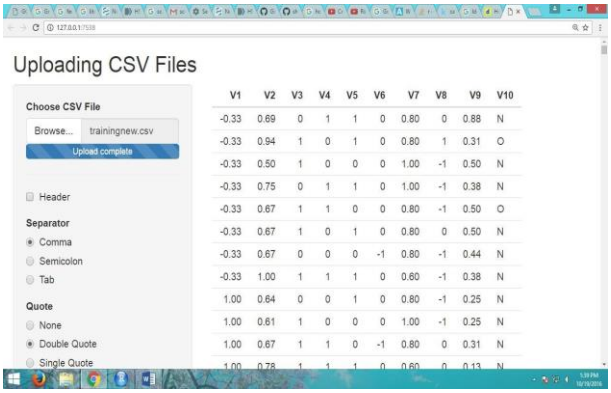

FIGURE 3. TRANING DATASET

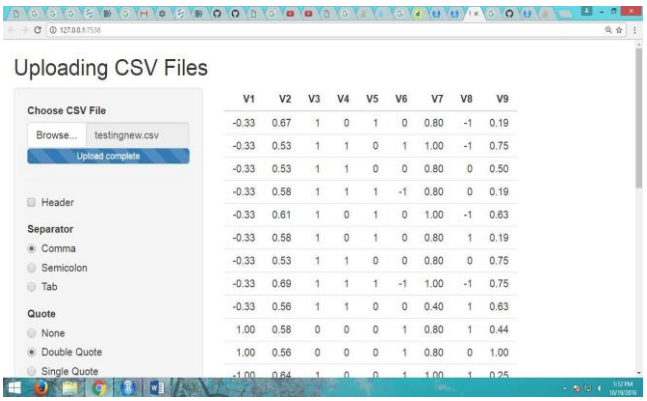

FIGURE 4. TESTNG DATASET

Random forest is applied with predictRF method, confusion Matrix and find results. By applying Latent Dirichlet allocation on same data with predictLDA metod , confusion Matrix and check out the results.

R Source code:

```
fitControlStruct<-trainControl(method="cv",number=10,
allowParallel=T, verbose=T)
randForestFit<-train(V10~.,data=trainingData,method="rf"
trControl=fitControlStruct, verbose=F)
predictRF<-predict(randForestFit, newdata=testData)
confusionMatrix(predictRF, testData$V10)
predictionValidation<-predict(randForestFit,  newdata=cleandstest)
predictionValidation
ldaFit<train(V10~.,data=trainingData,method="lda",trControl=fitCo
ntrolStruct,  verbose=F)
predictLDA <-predict(ldaFit, newdata=testData)
confusionMatrix(predictLDA, testData$V10)
```

V. RESULTS

Below is the result set found after implementing Random Forest and Latent Dirichlet Allocation algorithm.

TABLE III. COMPARISON OF RESULTS FOR RANDOM FOREST AND LATENT DIRICHLET ALLOCATION

Random Forest	**Latent Dirichlet allocation**
Confusion Matrix and Statistics Reference Prediction N O N 345 2 O 3 37	Confusion Matrix and Statistics Reference Prediction N O N 329 39 O 19 0
Accuracy : 0.9871	Accuracy : 0.8501
95% CI : (0.9701, 0.9958)	95% CI : (0.8106, 0.8842)
No Information Rate : 0.8992	No Information Rate : 0.8992
P-Value [Acc > NIR] : 1.972e-12	P-Value [Acc > NIR] : 0.9991
Kappa : 0.9295	Kappa : -0.0707

Mcnemar's Test P-Value : 1	Mcnemar's Test P-Value :0.0126
Sensitivity : 0.9914	Sensitivity : 0.9454
Specificity : 0.9487	Specificity : 0.0000
Pos Pred Value : 0.9942	Pos Pred Value : 0.8940
Neg Pred Value : 0.9250	Neg Pred Value : 0.0000
Prevalence : 0.8992	Prevalence :0.8992
Detection Rate : 0.8915	Detection Rate : 0.8501
Detection Prevalence : 0.8966	Detection Prevalence :0.9509
Balanced Accuracy : 0.9700	Balanced Accuracy :0.4727
'Positive' Class : N	'Positive' Class : N

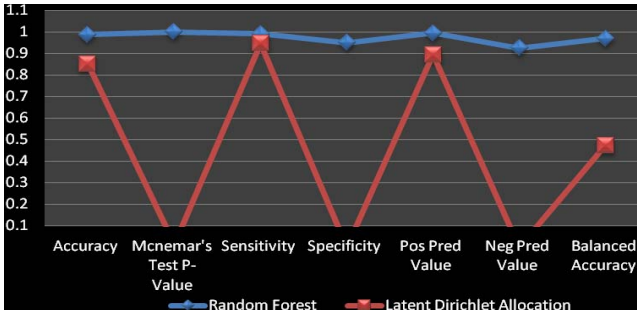

FIGURE 5. LINE GRAPH TO PORTRAY RANDOM FOREST AND LATENT DIRICHLET ALLOCATION RESULTS

From the experiment and above figure it can be concluded that Random Forest algorithm gives better accuracy.

VI. CONCLUSION

There are many different tools and techniques available using which big data can be analyzed wisely. R tool has been used for the analysis of data in combination with two different algorithms to come up with concrete result; case study has been performed wherein fertility data analyzed by applying Random Forest and Latent Dirichlet allocation algorithms in conjunction with R studio. Comparing results of case study suggests that Random Forest Algorithm has bit more accuracy than Latent Dirichlet allocation. As future scope of the research, soft computing techniques can be applied over big data to find out better accuracy.

REFERENCES

[1] Suthaharan, Shan. "Big data classification: Problems and challenges in network intrusion prediction with machine learning." ACM SIGMETRICS Performance Evaluation Review 41.4 (2014): 70-73.

[2] Jimeng Sun , Chandan K. Reddy , “Big Data Analytics For Healthcare” Tutorial presentation at the SIAM International Conference on Data Mining, Austin, TX, 2013.

[3] Priyanka P. Shinde, Kavita S. Oza, R. K. Kamat, “An Analysis of Data Mining Techniques in Aggregation with Real Time Dataset for the Prediction of Heart Disease”, in International Journal of Control Theory and Applications 9(20):327-336

[4] Amir Gandomi, Murtaza Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, Volume 35, Issue 2, April 2015, Pages 137-144, ISSN 0268-4012

[5] A. Nasridinov and Y. H. Park, " Visual Analytics for Big Data Using R," 2013 International Conference on Cloud and Green Computing, Karlsruhe, 2013, pp. 564-565. doi: 10.1109/CGC.2013.96

[6] Agnivesh and R. Pandey, "Elective Recommendation Support through K-Means Clustering Using R-Tool," *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur,2015,pp.851-856. doi: 10.1109/CICN.2015.324

[7] A. Malviya, A. Udhani and S. Soni, "R-tool: Data analytic framework for big data," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp. 1-5. doi:0.1109/CDAN.2016.7570960

[8] M. V. Datla, "Bench marking of classification algorithms: Decision Trees and Random Forests - a case study using R," 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15), Bangalore, India, 2015, pp. 1-7.doi: 10.1109/ITACT.2015.7492647

[9] S. d. Río, J. M. Benítez and F. Herrera, "Analysis of Data Preprocessing Increasing the Oversampling Ratio for Extremely Imbalanced Big Data Classification," *2015 IEEE Trustcom/BigDataSE/ISPA*, Helsinki, 2015, pp.180-185.doi: 10.1109/Trustcom.2015.579

[10] K. VijiyaKumar, V. Govindasamy and T. Esther, "An online big data take oution using latent dirichlet allocation," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, Tamilnadu, India, 2016, pp. 2278-2283. doi: 10.1109/ICCSP.2016.7754101

[11] David M Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3 (2003) 993-1022

[12] Open Government Data (OGD) Platform India[https://data.gov.in/]