International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 04, 2020
ISSN: 1475-7192

Performance Analysis of Naïve Bayes Correlation Models in Machine Learning

[1]Dr. K. Uma Pavan Kumar,[2]Dr. M. Kalimuthu

ABSTRACT--*Machine Learning(ML) usage ranges from individual research to developing predictive models for the prestigious organizations like reliance, Facebook, twitter and LinkedIn etc.,. The common point is usage of bulk data and applying some sort of the algorithms on that data and come up with predictive analytics or observing the useful patterns so as to reach the target customers and serve the public in a better way. The companies trying ML strategy to improve their business in drastic way and in many cases it has been proved.The current work focus on ML benefits and discussion of various algorithmic contexts like Naïve Bayes, Random Forest and Correlation analysis on certain data sets and our aim is to provide a basis of these algorithms and the usage models of algorithms along with some case studies. We believe that the work helps to understand the algorithms in simple way and helps the researchers to have some idea about the usage of the algorithms.To implement the algorithms R packages and methods we have used, R provides the importing the data and usage of the libraries related to algorithms and provides the plots so as to get the better understanding of the results.The significance of the work is describing the said algorithms along with research issues related to those aspects and publishing the results with analysis of the data sets. The outcome of the work is research issues related to the mentioned algorithms, result analysis and future scope of these works can be found.The algorithms naïve Bayes belongs to the category of supervised learning and comes under the category of classification techniques. Here supervised refers to the identified labels and expected outcome which can be achieved in the optimized way. The correlation analysis gives the idea about the kind of the relation between the entities which helps to keep track of the positive or negative correlation between the entities.*

Keywords-- *Algorithms, Naïve Bayes, R language, Correlation Analysis , predictive models.*

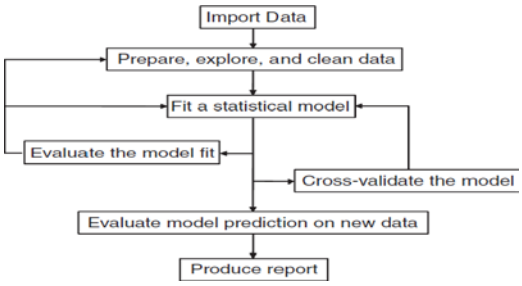
I. INTRODUCTION

In the analysis of the data with the help of ML algorithms the best possible languages are R/Python, in the current article the implementation presented is in the context of R. R provides flexible and most powerful packages and methods so as to implement the techniques of ML such as classification, clustering and recommender system kind of the activities.Before exploring the R implementation and discussion of the results every data scientist/analyst should know about the flow of the activities to be performed. In the analytics paradigm working with huge data sets is natural and better analytics are possible with huge amounts of the data only. So initially importing the data from the source might be HDFS/My SQL/RDBMS is mandatory. The data captured from the source might not be suitable to the algorithm, so pre-processing is needed so as to remove NULL values/missing values or any other kind of the unnecessary dimensions is the next step.

[1]*Associate Professor, Department of CSEMalla Reddy Institute of Technology, Hyderabad, India*

[2] *Associate Professor, Department of CSEMalla Reddy Institute of Technology, Hyderabad, India*

DOI: 10.37200/IJPR/V24I4/PR201088
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020



Figuer 1: Steps in the implementation of ML process

The ML algorithms are strongly dependent on the statistical methods so the corresponding statistical methods must be embedded. The next important step is to come up with a model which is the actual implementation of the algorithms based on the dataset. Here the point of training and test data comes to the picture, initially while building the model the training data can be given to the model and later the same model can be used to test with the help of test data set. Sometimes the strategy of with replacement is good otherwise without replacement can be used. The final step is to evaluate the model and conclude the results with proper inference.The organization of the article goes like this in section II the usage of Naïve Bayes has been mentioned, in Section III the usage of correlation analysis and the implementation has been mentioned, in Section IV the research issues and future scope is mentioned. In section V the conclusion has been mentioned.

II. USAGE OF NAÏVE BAYES AND IMPLEMENTATION WITH R

The life cycle of the analytics project is very important, before implementing any ML algorithm some steps need to be followed by every analyst.Initially we have to understand the business requirements, next the process of converting the business requirements into statistical problem. Once the statistical method is over, then selection of either R/Python so as to solve the statistical problem. The previous step gives the solution in terms of statistical measures, the same should be converted to business solution and the solution can be communicated to the client.The classification model specifies prediction of the unknown class labels; the categorical data such as Male/Female, blood groups comes under the category of categorical data if we make use of categorical data for analysis that can be termed as classification. Sometimes the data might be numerical in nature in that case the same process can be termed as prediction.

```
setwd("c:\\uma)
DB<-read.csv("Diabetes.csv",head=T) head(DB)
nrow(DB) set.seed(2)
DB$ind<-sample(2,nrow(DB),replace=TRUE,prob=c(0.7,0.3)) head(DB)
trainData<-DB[(DB$ind==1),] testData<-DB[(DB$ind==2),] nrow(trainData) nrow(testData)
library(e1071)
NB<-naiveBayes(Class.variable~Number.of.times.pregnant
+Plasma.glucose.concentration
+Diastolic.blood.pressure
+Triceps.skin.fold.thickness
+X2.Hour.serum.insulin
```

DOI: 10.37200/IJPR/V24I4/PR201088
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

+Body.mass.index

+Diabetes.pedigree.function

+Age..years., data=trainData) attributes(NB) NB$aprioriNB$tables

predNB<-predict(NB,testData,type=c("class")) ###Confusion Matrix#### table(testData$Class.variable,predNB)

head(trainData)

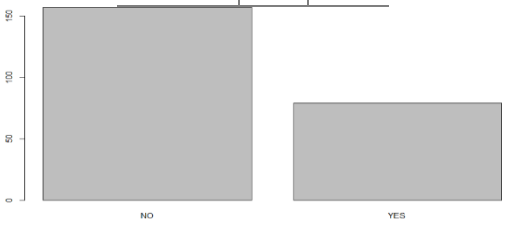
trainData1<-trainData[,-10] head(trainData1)

NB1<-naiveBayes(Class.variable~., data=trainData1)

predNB1<-predict(NB1,testData,type=c("class")) predNB1

plot(predNB1)

Results:		
predNB		
NO	NO	**YES**
	129	**23**
YES	**28**	**56**



Figuer 2: Prediction results from Naïve Bayes.

III. CORRELATION ANALYSIS

The Correlation analysis specifies the relation between various entities and with the help of P value the analyst can decide whether the entitieshave strong correlation or not. The P value is used to assess the NULL and alternative hypothesis. The P value 0.05 specifies the level of significance. P value is between 0 and 1, a small p-value indicates strong influence against the NULL hypothesis.

x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)

y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)

summary(x) summary(y) cor(x,y) boxplot(x,y) #Apply lm() relation<-lm(y~x)

print(summary(relation)) print(relation)

a<-data.frame(x=131) result<-predict(relation,a) print(result)

b<-data.frame(x=200) result<-predict(relation,b) print(result)

relation1<-lm(x~y) print(summary(relation1)) print(relation1)

c<-data.frame(y=61) result1<-predict(relation1,c) print(result1)

IV. RESULTS

summary(x)

Min. 1st Qu. Median Mean 3rd Qu. Max. 128.0 136.5 151.5 153.8 171.2 186.0

summary(y)

DOI: 10.37200/IJPR/V24I4/PR201088
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

Min. 1st Qu. Median Mean 3rd Qu. Max. 47.00 56.25 62.50 65.30 75.00 91.00

cor(x,y)

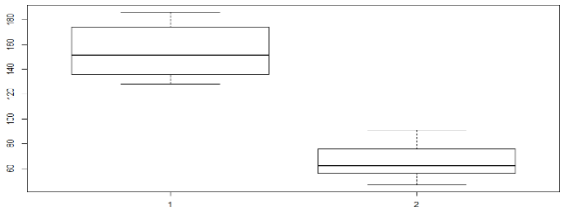
[1] 0.9771296

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) -38.45509 8.04901 -4.778 0.00139 **

x 0.67461 0.05191 12.997 1.16e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Figuer 3: Box Plot analysis of Correlation between the entities.

V. CONCLUSION

The work described the usage of Naïve Bayes algorithm and implementation with R programming. The results has been published and analysed. The correlation usage and implementation of the same with 2 variables and results has been described. The future scope is to implement the Naïve Bayes with some other revisions such as identification of high priority dimensions against to classifier, which helps to identify most influential dimension in the data set and according the classifier can be analysed.

REFERENCES

1. Kim Hazelwood, “Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective”, Facebook, 2017.
2. B. Reagen, et. al., “Deep Learning for Computer Architects, ser. Synthesis Lectures on Computer Architecture”, Morgan & Claypool Publishers, 2017.
3. J. M. Pino, A. Sidorov, and N. F. Ayan, “Transitioning entirely to neural machine translation”, Aug. 2017, https://fb.me/pino 2017.
4. U. P. K. Kethavarapu, “Various Computing models in Hadoop eco system along with the perspective of analytics using R and Machine learning”, International Journal of Computer Science and Information Security, vol. 14, pp. 17-23.
5. C. P. Chen and C. -Y. Zhang, “Data Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data”, Information Science, vol. 275, pp. 314-347, 2014.
6. J. Y. Monteith, J. D. McGregor and J. E. Ingram, “Hadoop and its Evolving Ecosystem”, 5th International Workshop on Software Eco System , pp. 57-68, 2013.
    a. S. Tanenbaum and M. Van Steen, Distributed systems. Prentice-Hall, 2007.
7. 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.
8. Shvachko, H. Kuang, S. Radia, and R.Chansler, “The hadoop distributed file system,” in 2010 IEEE

DOI: 10.37200/IJPR/V24I4/PR201088
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

26th symposium on mass storage systems and technologies (MSST), 2010, pp. 1–10.

9. M. K. Islam and A. Srinivasan, Apache Oozie: The Workflow Scheduler forHadoop. O'Reilly Media, Inc., 2015.
10. Bommareddy, m. &hebbar, . S. (2019) a review on pprom (preterm prelabour rupture of membranes) and early onset neonatal sepsis and role of inflammatory markers in diagnosis of maternal and neonatal infection. Journal of Critical Reviews, 6 (3), 7-13. doi:10.22159/jcr.2019v6i3.31792
11. Wei, X., Chen, X., He, L., Liu, L. Behavioral inhibition improvement through an emotional working memory (EWM) training intervention in children with attention deficit/hyperactivity disorder (2017) NeuroQuantology, 15 (2), pp. 261-268.
12. Gaiseanu, F. An information based model of consciousness fully explaining the mind normal/paranormal properties (2017) NeuroQuantology, 15 (2), pp. 132-140.

DOI: 10.37200/IJPR/V24I4/PR201088
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020