IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 40-43

Big Data Analytics of SVM and Naïve Bayes Algorithm for Multiple Datasets

Madhura A. Chinchmalatpure[1], Dr. Mahendra Dhore[2]
[1]*Department of Electronics and Computer Science, RTM Nagpur University Campus, Nagpur,(MS)-India*
[2]*Department of Electronics and Computer Science, RTM Nagpur University Campus, Nagpur,(MS)-India*

Abstract: *Big data is a collection of large datasets. Big data is structured, unstructured, semi-structured or heterogeneous in nature. For analyzing data, we used regression and machine learning as a statistical Technique. It shows the statistical relationship between two or more variables. The statistical technique can be evaluated for the predictive model based on the requirement of the data. This paper deals with two machine learning techniques Support vector machine and Neive Bayes applied on two databases, we create model using machine learning techniques which are compared using the training dataset in order to see correct model for better prediction and accuracy applied on database.*
Keywords: *Big Data Analysis, regression technique, machine learning technique*

I. Introduction

Big data is a collection of large datasets that cannot be managed efficiently by common database management systems.Big data is structured, unstructured, semistructured or heterogeneous in nature.In Digitized world, large amount of data is generated, to properly analyze that data big data concept is generated. Big data is thetermusedtodescribecollectionoflargeandcomplexdatasetshaving−4v‖definition.

volume(amount of data),
variety(range of data types and sources), velocity(speed of data in and out)
veracity (e.g.medical images, electronic Health Record (EHR), biometrics data etc.)

In healthcare industry large amount of data has generated so it uses Electronic Health Record(EHR) for patients data, clinical report, doctors prescription, diagnostic reports, medical images, pharmacy information, health insurance related data, data from social media and medical journals. All these information collectively forms big data in healthcare datasets Here we take two datasets as Antibiotics and medicare database is designed for clinical purposes.In this we have to compiles, summarizes, and organizes machine learning challenges with BigData.

Big data challenges can be divided into
1. Data Challenge: Volume, velocity, variety, veracity, Data Discovery
2. Processing Challenges: Data Collection, Modification of data, Data Analysis, outputrepresentation
3. Management Challenges: Data Privacy, Data Security, Governance and ethicalissues

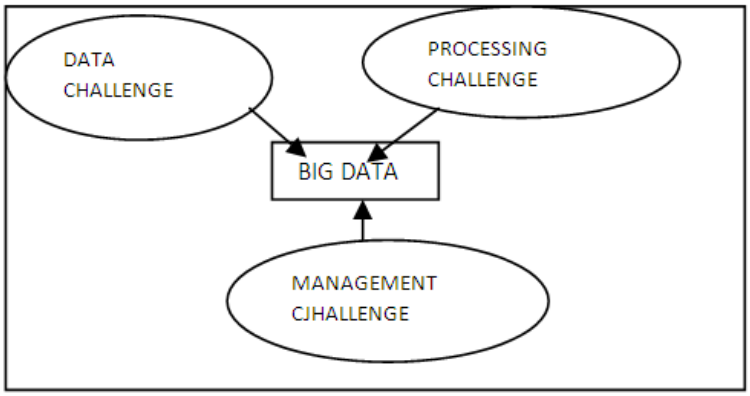

Fig*. Big Data Challenges*

Big data analytics is challenging research area, it refers to tool such as R that are applied to healthcare dataset to obtain the data from database in which to collect current data, preprocess data and analyze data.[1]
Analytics focus on statistical and mathematical analysis of data. The analysis helps to identify the problem from the collected data source. Later it uses various algorithms for better outcomes of data.
Machine learning is supervised machine learning is where we have input variables(x) and output variables(Y) and we use an algorithm to learn the mapping function from input to output

$$Y=f(X)$$

The goal of mapping function is when we have input data(X) we can predict the output variables(Y) for that data. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a supervising the learning process. Learning stops when the algorithm achieves an acceptable level of performance[2]

The machine learning techniques used in this research are applied on databases are as follows:

i) Neuralnetworks
ii) Multilayer perceptron(MLP)
iii) Radial basis functions
iv) Support vectormachines
v) NaïveBayes
vi) k-nearestneighbours

In this paper we are applying those techniques on two database as antibiotics and medicare databases. These benefits help data analysts to estimate the best set of variables to be used to build the predictive accuracies.

i) NeuralNetwork

In information technology (IT), a neural network is a system of hardware or software which is used for operation of neurons in the human brain. Neural networks -- also called artificial neural networks -- are a variety of deep learningtechnology, which also falls under the umbrella of artificial intelligence [3]

ii) MultilayerPerceptron(MLP)

A multilayer perceptron is a neural network in which connecting multiple layers with a directed graph, which means that the signal path through the nodes only goes one way. An MLP uses backpropagation as a supervised learning technique. Hence there are multiple layers of neurons, Multi-Layer Perceptron is a deep learning technique. In multilayer Perceptron, the most efficient models for data classification and prediction. The models are used to classify the variety of healthcare datasets.[5]

iii) Radial basisfunctions

Radial basis functions are means to approximate multivariable functions by linear combinations of terms based on a single univariate function. This is radialised so that in can be used in more than one dimension. They are usually applied to approximate functions or data which are only known at a finite number of points.It is explained that Radial basis Function Networks are used as the data classification for many real life problems with highest data classification accuracy.[5]

iv) Support Vector Machine

A support vector machine (SVM) is a supervised machine learning algorithm which is analyzes data for both classification and regression challenges. SVM is a supervised learning method that looks at data and sorts it into one of two categories and it analyzes the data and recognize patterns.
It has trained with a series of data already classified into two categories, building the model as it is initially trained. This is frontier which best segregates the two classes(hyperplane/line)[6]

v) NaïveBayes

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independent assumption. Naïve Bayes classifier produces probability estimates. [7]

vi) k-nearestneighbours

The *k*-nearest neighbours algorithm is one of the machine learning algorithms. It is simply based on the idea that‒objects that are‗ near‘ to each other will also have similar characteristics. Thus if you know the characteristic features of one of the objects, you can also predict it for its nearest neighbour.‖ k-NN is an improvisation over the nearest neighbour technique.[8] [9]

II. Proposed Algorithm

In this paper we discuss Support Vector Machine, Naïve Bayes techniques on two datasets Antibiotics, Medicare and calculate its accuracy and we discuss the algorithm of support vector machine and Neive Bayes are asfollows:

2.1 Support VectorMachine

A support vector machine (SVM) is supervised machine learning algorithm that used for both classification and regression challenges. This is a frontier which best segregates two classes for analyzes data.(hyperplane/line)

Algorithm: Generate SVM **Input**: Training Data **Output**: Decision Value

- Load dataset, Take out the experiment of gathering a sample of observedvalues
- Create a data frame ofdata
- Take out the training dataset of gather a sample of observedvalues
- Create a relationship model using the **lm()** functions inR.
- Fit a model. The function syntax is very similar to lmfunction
- For SVM(),we have to take the observed values and training data
- For prediction of data, take value of svm() and trainingdata.
- Use table(), and actual and predicted values arecalculated
- Take Matrices by using ConfusionMatrix(), and accuracy,Sensitivity,Specificity arecalculated

2.2 NaiveBayes

Neive Bayes is probabilistic classifier based on Bayes TheoremNeive Bayes computes conditional posterior probabilities of categorical class variables given independent predictor variables.

Algorithm: Generate Naïve bayes **Input**:Training Data **Output**:Decision Values

- Load dataset,Take out the experiment of gathering a sample of observedvalues
- Create a Partition ofdata
- Take training and testingdata
- Using naiveBayes() Create a model using trainingdata
- Take predicted values by using naivebayes() and testdata
- Use table(), and actual and predicted values arecalculated
- Take Matrices by using ConfusionMatrix(), and accuracy,Sensitivity,Specificity arecalculated
- Fitting the Neive Bayes model for discrete predictors, prior or conditional probabilities arecalculated
- The model creates the conditional probability for each feature separately. We also have the a-priori probabilities which indicates the distribution of ourdata.

III. Performance Evaluation

A. Classification of Datasets

To evaluate the effectiveness of our method, experiment on Antibiotics and Medicare datasets are conducted. They are publically available on the internet. Table shows the description of database.

Table: Description of Database

Sr.No.	Database	No. of Attributes	Size
1	Antibiotics	5	17
2	Medicare	17	10001

B. Comparative Analysis

Here we evaluate Accuracy of two databases using Support vector machine and Naïve bayes Technique The Value of SVM Technique in Antibiotics database is 0.6875 and in medicare dataset is 0.3611.
The Value of Naïve Bayes Technique in Antibiotics database is 0.5 and in medicare dataset is 0.9768.

S.No.	Algorithm	Datasets	Accuracy
1.	Support Vector machine	Antibiotics	0.6875
		Medicare	0.3611
2.	Naive Bayes	Antibiotics	0.5
		Medicare	0.9768

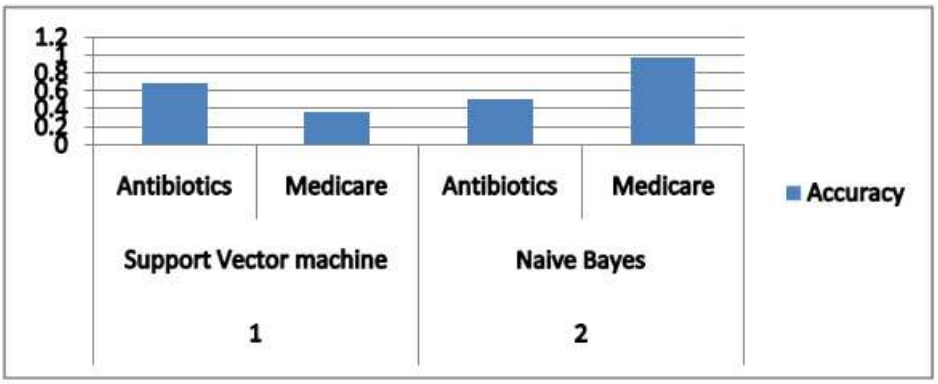

Fig: *Comparison between SVM and Naïve Bayes*

The performance of Naive Bayes shows high level compare with performance of SVM of ensemble classifier.

IV. Conclusion

In this paper, we examined the performance of the two machine learning techniques used for Big Data Analytics on three databases. We presented algorithms of these methods through regularized profile plots. The accuracy of classification technique is evaluated. The important challenge in machine learning area is to build efficient classifier for Medial Application. The performance of Naive Bayes shows high level compare with SVMof ensemble classifier. The confusion matrix of each classifier method is presented and the value of measure the performance of the method i.e. Accuracy is derived from confusion Matrix. It was found that naïve bayes model produced highest accuracy i.e. 0.9768 on Medicare dataset, which is so far highest. Other classifier like SVM were far less accurate compared to Naïve Bayes

References

[1]. Gemson Andrew Ebenezer J.1 and Durga S.2,‖ BIG DATA ANALYTICS IN HEALTHCARE: A SURVEY‖ ARPN Journal of Engineering and Applied Sciences ©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved. VOL. 10, NO. 8, MAY 2015, ISSN 1819-6608

[2]. Alexandra L'Heureux, Katarina Grolinger, Hany F. ElYamany, MiriamA. M. Capretz,-Machine Learning with Big Data: Challenges and Approaches‖, DOI 10.1109/ACCESS.2017.2696365, IEEE Access

[3]. Guoqiang Peter Zhang, -Neural Networks for Classification: A Survey‖, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 30, NO. 4, NOVEMBER 2000

[4]. Wilbert Sibanda, Philip Pretorius,‖Artificial Neural Networks- A Review of Applications of Neural Networks in the Modeling of HIV Epidemic‖, International Journal of Computer Applications (0975 – 8887) Volume 44– No16, April2012

[5]. Shrwan Ram Dr. N.C. Barwar,‖A Comparative Study of Multilayer Perceptron, Radial Basis Function Networks and logistic Regression for Healthcare Data Classification‖, Volume 3, Issue 3, March-2016

[6]. ShashikantGhumbre, Chetan Patil, and Ashok Ghatol,‖Heart Disease Diagnosis using Support Vector Machine‖, International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya Dec. 2011

[7]. JanmenjoyNayak, BighnarajNaik* and H. S. Behera ,‖A comprehensive survey on support vector machine in data mining tasks: Applications & challenge‖, International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.169-186 http://dx.doi.org/10.14257/ijdta.2015.8.1.18

[8]. Dr. S. Vijayarani1, Mr.S.Dhayanand,‖ Liver Disease Prediction using SVM and Naïve Bayes Algorithms‖, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015

[9]. Hassan SheeKhamis, Kipruto W. Cheruiyot, Stephen Kimani ,‖Application of k-Nearest Neighbour Classification in Medical Data Mining‖,International Journal of Information and Communication Technology Research ,Volume 4 No. 4, April 2014 ISSN 2223-4985

[10]. Kathija,ShajunNisha,‖breast Cancer Data Classification Using SVM and Naïve Bayes Technique‖,International Journal of Innovative research in Computer and Communication Engineering,Vol.4,Issue 12,December2016,ISSN(online):2320-9801

[11]. Preety,SunnyDahiya,‖Sentiment Analysis Using SVM and Naïve Bayes Algorithm‖,International Journal Of Computer Science and Mobile Computing,IJCSMC,Vol. 4,Issue 9,September 2015,ISSN 2320-088X