# Study to Analyze the Performance of Supervised Machine Learning Classifiers

**S.Magesh, G.Chandar**

***Abstract: On the rise exponentially, machine learning is adopted by various organizations across various domains to crunch petabytes of data to deliver more meaningful conclusions and predictions. With machine learning, it's become quite possible to transform a wide variety of industries, in this complicated world. A category of machine learning algorithms called as classifiers, have found wide spread use in processing the large datasets efficiently. Classifiers are applied to big datasets through supervised learning methods. Though there exist huge number of classifier algorithms, it is important to apply some intelligence in careful selection of algorithms, to deliver excellent business intelligence. In order to select a more suitable classifier algorithm for a problem, it is important to analyze the performance of classifiers based on certain parameters and different datasets. In this paper, we analyzed variety of classifiers from the perspective of performance measures such as Precision, Recall, Mean Absolute Error, and Root Mean Square Error pertaining to two different datasets using WEKA. Also, in the conclusion we present the findings from analysis of performance of classifiers.***

***Keywords:Classifiers, Supervised , Machine learning, Precision, Recall,WEKA***

## I. INTRODUCTION

Machine learning is technique in which a machine or system is trained to learn and perform based on the training sets of data for the required task. The success of machine learning by a system depends upon the accuracy and precision in which the system in subject processes the input data based on the training it had obtained. Machine learning is a process through which intelligence is built into the system in an artificial manner. One of the familiar examples of machine learning includes filtering of email spam and their categorization by email service provider.[9]

The entire concepts of machine learning can be divided into Supervised learning, Unsupervised learning, Reinforcement learning, Semi supervised learning and deep learning. Under supervised learning, systems are provided with training inputs and desired outputs, through which a mapping between input and output is obtained by the system, for further automated processing of inputs. In unsupervised learning, no training labels are provided to the system, leaving the system on its own to find a structure in the data.

Under supervised learning, most predominantly used techniques are artificial neural network, Bayesian statistics, and classifiers. In supervised learning based classification, the training data are accompanied by labels indicating the class of observations. New data is classified based on the training set. The major applications of classification algorithms under supervised learning are credit approval, target marketing, fraud detection, medical diagnosis etc. Social media plays a major role in fetching huge datasets which shall be used in variety of applications of machine learning classifications to provide excellent business intelligence for various organizations[8]. Normally classification is considered to be a two step process where construction of a classifier model and usage of constructed model are involved[1][2]. Choosing an appropriate classification algorithm based on machine learning is really important for deriving meaningful and efficient results from the provided input datasets. The biggest bet with respect to the selection of classification algorithm lies with the accuracy of classification task. A simple solution is to try out different algorithms with the datasets and choose the best algorithm by cross validation. The size of the data set matters most when it comes to the performance of any classification algorithm.[6][7]. For a smaller dataset, high bias/ low variance classifiers are referred over low bias/high variance classifiers. As training set size grows, the preference will be given to low bias/high variance classifiers. Best in class examples for high bias/low variance and low bias/high variance classifiers are naive bayes and KNN respectively.[3][18]

A different perspective in generating a classification model lies in creating a generative model or discriminative model. A generative model utilizes joint probability distribution and a discriminative model utilizes conditional probability distribution. Literatures prove that for many classifications tasks discriminative models outperform generative models[4][18]. Both the above mentioned models fall into the class of supervised learning. Generally the practitioners of machine learning preach the concept of not to learn anything that is irrelevant and extra from what is present. This basic practitioner's ideology itself is defeated during the creation of generative models. The adoption of a various classification algorithms for various tasks can be selected from the bunch of popular algorithms such as linear classifiers and Decision trees[5]. This paper will analyze the performance of classifiers such as Logistic regression, Naive bayes, Support Vector Machines and multi layer perceptron under linear classifiers, J48, Random forest, Hoeffding Tree under tree based classifiers. The next section of this paper focuses on the descriptions of the above mentioned classification algorithms with specific relevance to applications of classifier algorithm and suitability of

**Revised Manuscript Received on April 05, 2019.**

**S. Magesh,** Professor, Department of Computer Science, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (mageshs.sse@saveetha.com)

**G. Chandar,** Research Scholar, Department of Computer Science, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (chan.it13@gmail.com)

larger/streaming datasets. Section II provides the basic review of popular classifiers that supports for analytical methods.Section III deals with the functionalities and capabilities of WEKA along with the description of the representative datasets used in analyzing the performance of various classifiers in this paper. Performance measure analysis of various algorithms and conclusion and future work are dealt with in sections IV and V respectively.

## II. RELATED WORKS AND REVIEW OF POPULAR CLASSIFIERS

Inductive Learning principle is considered to be the basis for development of various classifiers. According to this principle, any hypothesis that can approximate a target function over large training sets will approximate the same target function well over other unobserved examples. A set of training observations $(x_1,y_1)$, $(x_2,y_2)$,...$(x_n,y_n)$ are used to build a classifier. Typical categories of classification algorithms include rule based, tree based, function based, neural networks and genetic algorithms. In this paper, we present the performance analysis of various classifiers under the hood of linear classifiers and tree based classifiers. In this section, relevant descriptions of the classification algorithms with their application domains and suitability for large and streaming datasets will be provided.

Logistic regression is a classification technique that will compute the conditional probability of binary output variable Y as a function of the input variable X. Maximum Likelihood Estimation(MLE) is used in logistic regression to estimate any unknown parameters in the function. The application of logistic regression dates back to 1845 when this technique is used to analyze the population growth based on mathematical techniques. Logit transformation is applied to the dependent variable and hence the term logistic regression is coined (Hair, Black, Babin, Anderson and Tahtam, 2006)[15]. Since MLE is used in logistic regression instead of least squares principle, number of observations in the analysis should be more as it increases the reliability of the model[14][16]. Logistic regression is applied in all areas where the correlation of the dependent and independent variables are to be found and prediction has to be done based on the correlation.(Omaycokluk , 2010).

Naïve Bayes classifier is used in various applications in everyday activities such as Email spam identification, News article classification into technology, sports, politics, etc, face recognition softwares and identification of texts expressing positive or negative emotions. Naive bayes classification is relatively faster compared to other classification . The mathematics behind the Naïve Bayes classification algorithm is Bayes theorem of probability, assuming conditional independence among the input attributes. Naïve Bayes classification provides the best classification class for the target variable based on the training phase and classification phase. The mathematics governing the operation of Naïve Bayes classifier is given below:

$$P(H/X) = \frac{P(X\backslash H)P(H)}{P(X)}$$

Apart from Support Vector Machine (SVM) playing a vital role in statistical learning theory,[17] it has widespread application in Bioinformatics, face recognition, image processing and text mining. (Lipo Wang et al, 2005). SVM is a discriminative classifier formally defined by a separating hyperplane Givenlabeled training data, SVM optimally classifies which hyperplane the new examples should be categorized. The operation of SVM lies on identifying the largest minimum distance of the new examples in the hyperplane to the training examples. This distance is known as the margin in SVM's theory.[15] Normally hyperplanes are defined using the example

$$f(x) = \beta_0 + \beta^T x,$$

Where $\beta_0$ is known as the bias and $\beta^T$ is known as the weight vector. The training examples that are closest to the hyper plane are termed as Support Vectors.

Decision trees are a type of recursive partitioning algorithm. Decision trees are built up of two types of nodes: decision nodes, and leaves. The decision tree starts with a node called the root. If the root is a leaf then the decision tree is trivial or degenerate and the same classification is made for all data. For decision nodes we examine a single variable and move to another node based on the outcome of a comparison. The recursion is repeated until we reach a leaf node.

Different implementations and variations of the decision tree algorithm, such as Random Forest, J48 method which is a Java implementation of the C4.5 algorithm

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one.[19][20].

## III. EXPERIMENTAL SETUP AND DATASET

Weka3.7.13 supports many classifiers under classification [10]. Few categories of classifiers are enlisted under Bayes,Functions,Rules, Tree, Meta etc.[11]The dataset used for the proposed work is the weather dataset of Chennai for two years 2015 and 2016. The dataset contains 365 days of weather parameters such as temperature,humidity,rainfall, wind speed etc. Courtesy: https://www.wunderground.com/in/chennai. The wunderground.com site collects the weather data for the location and helps in forecasting the data generated by the National Weather Service's National Digital Forecast Database (NDFD). The weather data contains information related to IST temperature, Dew Point, Humidity, Sea Level Pressure, Visibility, Wind Speed, Gust Speed, Cloud Cover, Precipitation, Events, and WindDirDegrees etc.[13]

## IV. PERFORMANCE ANALYSIS AND RESULTS OF VARIOUS CLASSIFIERS USING WEKA

In the previous sections of this paper we presented with the intricacies of various machine learning techniques. In continuation with the same, this section will focus on the performance analysis of various classifiers such as Naïve

Bayes, BayesNet, Simple Logistics, Multi Layer Perceptron and j48. This analysis has been carried with huge instances of weather data sets obtained from Face book. The tool we used to carry out analysis is WEKA [12]and the metrics of precision, recall, Mean Squared Error and Root Mean Squared Error were used to present the performance of various classifiers.[11] The*Table .1* provides the details of the performance of various classifiers with respect to the above mentioned metrics.

Based on the analysis and observations from the *Table .1* and *Table .2* j48 classifier outperformed all the other classifiers in correctly classifying the number of instances followed by Naïve Bayes, BayesNet, Simple Logistics, and Multi Layer Perceptron. Also, the j48 classifier obtained the highest precision in classification relative to other classifiers.

In order to study the behaviour of classifiers, we provided the sample datasets belonging to the Weather dataset from [19] in WEKA to analyze the performance of classifiers.

The *Figure .1* and *Figure .2* proves that J48 classifier outperforms other classifier in terms of precision, recall and the time taken to build the model is also comparatively moderate when compared to Multi-Layer Perceptron ,logistic regression as tabulated.

| Classifier | P | R | MAE | RMSE | Time |
|---|---|---|---|---|---|
| Naïve Bayes | 0.70 | 0.61 | 0.10 | 0.29 | 0 |
| Logistic Regression | 0.72 | 0.72 | 0.08 | 0.24 | 0.31 |
| Decision Table | 0.68 | 0.75 | 0.11 | 0.22 | 0.09 |
| J48 | 0.70 | 0.73 | 0.08 | 0.24 | 0.09 |
| Random Tree | 0.61 | 0.69 | 0.11 | 0.24 | 0 |
| Multi layer Perceptron | 0.54 | 0.57 | 0.11 | 0.31 | 321.2 |

**Table .1 Performance of classifiers for Chennai Weather_2014**

| Classifier | P | R | MAE | RMSE | Time |
|---|---|---|---|---|---|
| Naïve Bayes | 0.68 | 0.54 | 0.13 | 0.34 | 0 |
| Logistic Regression | 0.70 | 0.69 | 0.10 | 0.26 | 0.22 |
| Decision Table | 0.64 | 0.70 | 0.14 | 0.25 | 0.07 |
| J48 | 0.72 | 0.75 | 0.09 | 0.26 | 0.02 |
| Random Tree | 0.65 | 0.68 | 0.13 | 0.27 | 0.01 |
| Multi layer Perceptron | 0.62 | 0.55 | 0.13 | 0.35 | 324.65 |

**Table .2Performance of classifiers for Chennai Weather_2015**

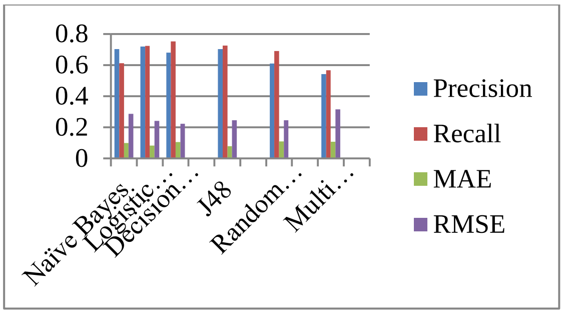

**Figure .1Performance evaluation measures of classifiers for Chennai Weather_2014**

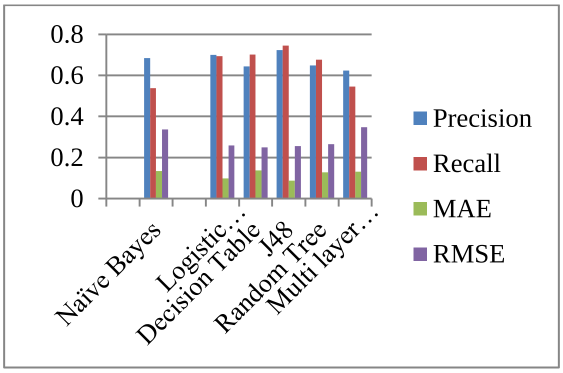

**Figure .2 Performance evaluation measures of classifiers for Chennai Weather_2015**

## V. CONCLUSION AND FUTURE WORK

The paper aimed to analyse the performance of machine learning classifier taking into account the weather dataset obtained from Wunderground. The work clearly concludes that the performance of any machine learning classifier totally depend on the nature of the data attributes and its size. Larger the size of the data sets then improved classification techniques should be utilized to obtain the desired measure. Future work could focus on the application of advanced classification techniques on Big data holding 10 to 20 years of weather data to forecast and predict future years weather.

## REFERENCE

1. J. PradeepKandhasamy, S. Balamurali, “Performance Analysis of Classifier Models to Predict Diabetes Mellitus “,Procedia Computer Science 47 ( 2015 ) 45 – 51
2. NongyaoNai-arun, RungruttikarnMoungmai, “Comparison of classifier for the risk of Diabetes Prediction “,Procedia Computer Science 69 ( 2015 ) 132 – 142
3. P. Shanthakumar[a,] ’, P. Ganeshkumar[b], “Performance analysis of classifier for brain tumor detection and diagnosis “,Computers and Electrical Engineering 45 (2015) 302- 311.
4. KarthikeyaniV,Parvin Begum I,TajudinK,ShahinaBegam, “ Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction”International Journal of Computer Applications, 2012; 60:26-31.
5. AmitGupte, Sourabh Joshi, Pratik Gadgul,AkshayKadamAmitGupte et al, “Comparative Study

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

of Classification Algorithms used in Sentiment Analysis", (IJCSIT)International Journal ofComputer Science and InformationTechnologies,Vol.5 (5), 2014,6261-6264

6. https://classeval.wordpress.com/literature-analysis/ Literature analysis on classifier evaluation.
7. Jose A.Lozano, Guzman Santafe, InakiInza , "Classifier performance evaluation and comparison",International Conference on Machine Learning and Applications (ICMLA 2010) December 12-14, 2010.
8. Martin Weis1, Till Rumpf2, Roland Gerhards1, Lutz Plümer1, "Comparison of different classification algorithms for weed detection from images based on shape parameters"
9. Vijayarani,J. Ilamathi,Nithya, " Preprocessing Techniques for Text Mining - An Overview" *International Journal of Computer Science & Communication Net-works,Vol 5(1),7-16 ISSN:2249-5789*
10. http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf
11. WEKA Approach for comparative study of classification algorithm by Trilokchand sharma[1],Manoj Jain[2] International journal of Advanced Research in Computer and communication Engineering vol.2,Issue 4,April 2013.
12. Evaluation of various classification Techniques of WEKA using different Datasets by Ramesh, Prasad ,Aharwal IJARITE-ISSN(O)2395-4396 vol -2 Issue -2 2016.
13. http://www.weatherzone.com.au/facebook-twitter/
14. ÖmayÇOKLUK,Gurcan (1998)," Logistic regression and its Applications, EDUCATIONAL SCIENCES: THEORY & PRACTICE
15. Hair, J., Black, B. Babin, B., Anderson, R. and Tatham, R. (2006) Multivariate Data Analysis(6thedition). Upper Saddle River, NJ: Prentice-Hall.
16. Cokluk, Omay, "Logistic Regression: Concept and Application Educational Sciences: Theory and Practice, v10 n3 p1397-1407 Sum 2010
17. Lipo Wang, "Support Vector Machines: Theory and Applications, Springer, Berlin, Germany (2005)
18. Magesh.S.,Nimala.K., "A survey on Machine Learning Approaches to Social Media Analytics" Volume 11,Number 4 pp 2411-2416 , (2016)
19. https://www.wunderground.com/in/chennai
20. https://en.wikipedia.org/wiki/Multilayer_perceptron