International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Improving the Performance Metrics of Binary Classification Models with Different Transforms

K. Issac[1], S. Abhinav Karthik[2], K. Sahana[3]

[1]*Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Chennai-India-600062*
[2]*Industrial Management, University of central Missouri, 116 W South St, Warrensburg, MO 64093, USA*
[3]*Master's in Computers and Information Sciences, Ohio, USA.*

Abstract - *Machine Learning became a part of our daily life. Google Maps, Face Recognition in social applications, Voice assistants and many more which helps us a lot. These are made under machine learning under different Categories. Where Supervised learning is one of the types which we used in the paper. The paper is an implementation of different Algorithms like KNN, RFC, NB, SVM, Logistic Regression with different Transforms and ensemble methods to improve the performance metrics like Accuracy using the Sonar Dataset.*

Key Words: Machine Learning, Supervised Learning, KNN, RFC, NB, SVM, Logistic Regression.

1. INTRODUCTION

Machine learning and Artificial Intelligence is the word which is being ringing in our ears in recent times. Machine learning is a subset of the Artificial Intelligence where the machines are being trained in to work with itself with being programmed separately or explicitly. Machine learning is seen around us where we can consider Tela smart cars, Google Maps, Social Application Tagging, personal Voice Assistants and many more… To implement or create these types of projects, where everything is of its own version or model. Machine Learning needs to be classified in four different types:

- Supervised Learning
- Unsupervised Learning
- Semi Supervised Learning
- Reinforcement Learning

In this Paper we consider Supervised Machine Learning which can be defined as training the machines under the supervision of a user or teacher or programmer. So, the data given for training the supervised Classifier Model will be having Features and labels for every feature, where the machine has to find out the relationship between Features and Labels and Logic output what we get is been tested. Consider X as the number of features and Y as the labels for these features then the logic is represented as the function of X to the Y as demonstrated as in equation 1.[1]

$$Y = f(X) \quad (1)$$

Where f is the logic to be tested and that logic will be tested to get the performance metrics. In this paper gives an idea on How classifiers work and how they enhance the performance metrics and next sections will be having the details of Introduction to the dataset, Basics of Classifiers and Transforms, Implementation and Methodology, Analysis and results, Conclusion and references.

2. INTRODUCTION TO THE DATASET

Sonar Dataset which had the title 'Sonar, Mines vs Rocks' used by Gorman and Sejnowski in their research about classification and neural networks. The Data is collected based on sending the sonar signals to classify between rocks and minerals. The csv (comma separated values) file contains 111 patterns are metal cylinders where 97 are came from rock, in the grand total there are 208 instances and every pattern has 60 numbers from 0.0 to 0.1 and 61st column indicated the label or target as 'R' or 'M'. Figure 1 depicts the histogram of the Dataset.

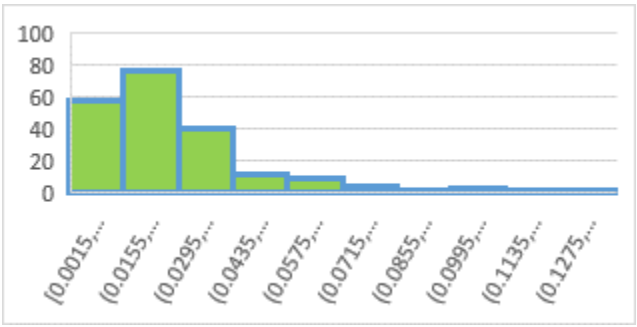

Figure -1: Histogram of sonar Dataset depicting the values ranging

Since all the 208 instances are ranging from 0.0 to 1 there is no need to perform normalization or any preprocessing techniques. The Dataset now to split into two parts of the ration 3:1 so that 75 percent of the data is used for the training the model and 25 percent is used for testing the model and function stratified shuffle split is used splitting the data in shuffled manner where it will increase performance since all the rock values are in ordered manner, this function can be imported from the scikit learn library which is a vast library containing all machine learning algorithms. The next session includes the basics of the

classifiers and transforms and the process involved to get the performance of model [2].

3. BASICS OF CLASSIFIERS AND TRANSFORMS, ENSEMBLE METHOD

Supervised algorithms are the models with trainers and they take the labelled data, produce the logical output. The calculation of that output is based on different parameters like KNN classification done based on the Distance formulation. Where the transforms used can be applied in Image processing domain representation etc….[3]

3.1 K-Nearest Neighbors

It is a non-parametric Algorithm and also called lazy learners' algorithm because it doesn't learn from the training data directly, it only learns or classifies when a new data point is on board. Then user has to give the K as the neighbors and it calculates the distances between neighbors and new point and based on the closest distance it classifies. In our case as 'R' or 'M'

3.2 Logistic Regression

Parameter or the working principle to be considered here is Probability, Logistic regression calculates probability of the class to be happen. There different types of Logistic Regression based on output. The sigmoid function acts as the activation layer where line is feed.

Types of Logistic Regression:

1) Binomial Logistic Regression
2) Multinomial Logistic Regression
3) Ordinal Logistic Regression

3.3 Support Vector Machines

SVM deals with the planes and graphs in 3 dimensional plots. Most important term is hyperplanes, based on the creation of the hyperplane classification depends. SVM acts as both Regressor and Classifier under supervised learning. Identifying the Hyper plane and Placing the Margin are the major roles in SVM.

3.4 Random Forest Classifier

It is a combination of many decision Trees and to defined it an ensemble method of classification and Regression. So it works based on the working of Decision Trees and Leaf nodes will be classification outputs and multiple decision Trees will help more iteration which leads to accurate results.

3.5 Navie Bayes

From the word we can understand that naïve Bayes algorithm works on the principle of Bayes theorem. Where naïve, this word leads to the meaning that occurring of an event without another event intervention. Equation 2 represents the Bayes theorem.

$$P(A|B) = (P(B|A).P(A))/P(B) \quad (2)$$

3.6 Hilbert Transform

When you relate the image processing domain and machine learning domain, Hilbert transform places a obtaining minimal phase response from the spectral analysis, that means it identify the instant frequency changes.[4]

3.7 Discrete wavelet Transform

With a series of order functions known as wavelets which means small waves, each with a separate scale and localised with time. Discrete wavelet transform is used to get the octave scale and spatial timing of the signal.[5]

3.8 Principal Component Analysis

Transforming the large set of dimensions to smaller ones is what PCA does. Which is called as Dimensionality Reduction. The number of components the user gives as the input as one of its parameters it processes the data.[6]

3.9 Ensemble Method

The process of creating multiple models and feeding them into a list and iterating over them and performing some useful techniques such as voting, bagging, averaging, to get the more accuracy is called Ensemble Method. [7]

4. IMPLEMENTATION AND METHODOLOGY

For Every Process to begin, there requires a flow control or process map to followed to not make faults during the Flow of research. Figure 2 represent the Flowchart of this research done:

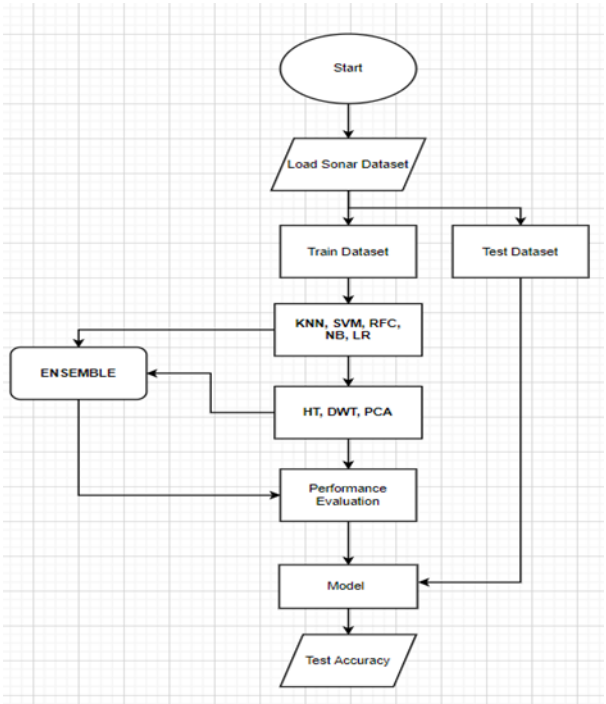

Figure-2: Flowchart of process to be followed

• Process Starts with Loading the Dataset as csv file using pandas data frame, pandas library helps the data in any file format to look good by using different functions. Then data is being split into two unequal parts in the ratio of 3:1 where 3 parts of data is used for Training and 1 part for Testing Data.

• Now for the training data we will be performing some operations with supervised algorithms and transforms to check which combination of algorithm and Transform yield more accuracy.

• Feed the Data into the transforms, Hilbert Transform (HT), Discrete Wavelet Transform (DWT), Principal Component Analysis (PCA) Separately.

• Feed the Data in Algorithms KNN, Support Vector Machines, Naïve Bayes, Logistic Regression, Random Forest Classifier Separately.

• Now for each Classifier and each transform combination and with no transform combination calculate the Performance Metrics, Accuracy, Mathews Correlation Coefficient, Balanced Accuracy.

• Simultaneously, Feed the five algorithms in ensemble method, Now calculates performance metrics for ensemble method for three transforms.

• There will total 24 Combinations including of all algorithms and transforms and ensemble methods.

• The next section continues with results and Discussion.

5. Results and Discussion

The Results and Discussion includes the information regarding the performance of the model with different metrics like Mathew's correlation coefficient, Accuracy, balanced accuracy, Error rate, F1_score, Recall, and Precision, these are the metrics we can access the scikit learn library. And to calculate these metrics manually with a confusion Matrix where it consists of True positives and True Negatives, False positives and False Negatives with different combinations of these TP, TN, FP, FN we can get these metrics. The following tables shows the MCC, Accuracy, Balanced Accuracy.

The table 1 shows the Accuracy values of the different values of Classifiers and transforms where SVM and RFC with HT gives 92.30.

Classifiers/ transforms	KNN	RFC	SVM	NB	LR	Ensemble
Without transform	88.46	86.53	90.38	59.61	82.69	86.53
HT	84.615	92.30	92.30	78.8	80.76	78.84
DWT	76.92	73.07	84.6	71.15	59.61	75.0
PCA	65.38	63.4	73.07	69.2	63.4	75.0

Table-1: Accuracy Values of the model.

Classifiers /transforms	KNN	RFC	SVM	NB	LR	Ensemble
Without transform	78.5	72.89	80.6	29.06	65.92	73.94
HT	72.0	84.8	84.52	62.23	61.3	57.3
DWT	54.26	45.83	69.53	43.61	17.68	49.59
PCA	30.2	26.288	46.93	37.81	26.28	49.59

Table-2: MCC values of Model.

Classifiers/ transforms	KNN	RFC	SVM	NB	LR	Ensemble
Without transform	87.5	86.30	90.178	61.90	83.03	85.71
HT	83.333	92.5	92.26	80.05	80.65	78.57
DWT	75.89	72.9	83.92	71.72	58.03	74.40
PCA	63.98	63.09	71.72	68.45	63.09	74.40

Table-3: Balanced Accuracy of the Model.

The Table 2 and 3 tells the balanced accuracy and Mathew's Correlation coefficient have the maximum of 92.5 for RFC with HT and maximum of 84.8 with RFC with HT. From these results, we can conclude that a machine learning model of Support vector machine and Random Forest classifier with Hilbert transform will be very helpful for training a data like sonar.

ACKOWLEDGEMENT

We Acknowledge here by our thanks to the sources we referred through and books we gone through. And past works related to all these works been so much helpful to this research work. And we forward our thanks to everyone who can't be mentioned here.

REFERENCES

1) Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J.O., Olakanmi O. and Akinjobi J. 2017 Supervised Machine Learning Algorithms: Classification and Comparison International Journal of Computer Trends and Technology 48 128-138J.

2) Bharanidharan N. 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC) (Depok, Indonesia) Harikumar Rajaguru: Dementia MRI Classification Using Hybrid Dragonfly Based Support Vector Machine.

3) K. Issac *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* 1084 012032.

4) King and Frederick W. 2010 Hilbert Transforms (Cambridge University Press) 1.

5) Sundararajan D. Discrete Wavelet Transform: A Signal Processing Approach (Wiley).

6) Olliffe I.T. 2002 Principal Component Analysis, Series (Springer Series in Statistics).

7) Shalinee Chaurasia and Anurag Jain. 2014. Ensemble neural network and k-NN classifiers for intrusion detection. International Journal of Computer Science and Information Technology 5 (2014), 2481—2485.