ACTA SCIENTIAE

ISSN: 2178-7727
DOI: 10.17648/acta.scientiae.6369

INSIGHTS OF FEATURE SELECTION IN IDS USING THE HYBRID PCALDA ALGORITHM

A Vinitha[1] , Dr. B. Rosiline Jeetha[2]

[1] Doctoral Research Scholar, Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore, vinithasmsc@gmail.com

[2] Doctoral Research Supervisor, Associate Professor & Head, Department of Computer Science, Nirmala College for Women, Coimbatore, jeethasekar@gmail.com

Abstract

The IDS is the very essential need for the digital communication to secure the data and to classify the connections as the normal or abnormal. The primary objective of the intrusion detection is to classify the connection. It is used to predict the model. There are various machine learning algorithms which will be used to predict the connection as the normal or abnormal. Machine learning algorithms are very useful in finding the solution for the difficult and challenging problem. In this various meta classifier algorithms are chosen and among them the best techniques are used for feature selection process in IDS.

Keywords: PCALDA,

Introduction

There are various machine learning algorithms. Among them the best one which is suited for the IDS feature selection is chosen. The IDS is very useful for the organization to predict the intrusions that occur inside and the outside of the organization. The main work of IDS to classify it as normal or malicious. There are many possibilities of occurring threat which will result in loss of information. There are different types of attacks that occurs in the network. The feature selection is the very important strategy for predicting the packet it as normal or intrusion. There are different machine algorithms which will help to predict the attacks. The author uses different types of features for selecting the features in the data set. The data set is the one which should be known in detail and to understand the importance of the attribute whether it impacts the result or not.

The meta classifiers like Adaboost, Logit boost, Vote, Stacking and iterative optimizer are the different meta classifier which is compared in this study to carry out the research. Initially all methods will undergo the training phase and later on the results will be compared. The data set used in this study is fully based on the LAN network. The data set contains the so much of packets which contains the information regarding to the source and the destination ports. There are 41 features among that 3 are qualitative and 38 are quantitative attributes. The qualitative data is some thing which is to be measured in terms of range. The quantitative data is one which can be exactly mentioned in numbers. In the data set some features are continuous and others are discrete. The discrete values can be represented in terms of count. The continuous values can be represented in terms of measurable amounts.

Literature survey

The author [1] et al states that the intrusions will occur in the digital era should be handled properly to secure the data. Existing models will help the user to optimize the feature selection

process. It mainly focuses on correlation between the data items which helps to improve the performance. The Pearson correlation algorithm is used to find the correlation.

The author [2] et al proposes that the network data are very complex and large in size. Many features need to be analysed to improve the performance of the system. The chaotic crow search algorithm is one which helps to find the features for binary and multiclass classifications. Many different classifiers are used to classify the data. The results shown the best one which will be very useful in increasing the performance of the system.

The author [3] et al tells that different algorithms are used for classification of attacks. Correlation based feature selection methods are used. It is an efficient method for analysing the intrusion and to secure the data.

The author [4] states that the intrusion prediction is the very difficult task to classify the patterns. Many IDS is prone to false positives. In this paper wrapper-based feature selection is used. In that correlation-based feature selection using best first search is used for comparison. The features are reduced from 78 to 4 features.

There was the more difficult in maintaining the network system because of the complexity of the network [5]. The author has used the anova f value-based method, impurity-based selection method, mutual information-based techniques are used for identifying the best features to detect the intrusion.

The machine learning based methods are chosen for finding the solution for the cyber-crime. There should be generic algorithm that decides the higher accuracy for the algorithm [6]. The main goal of enterprise is to protect the data.

Due to the usage of IOT devices in more there are more chances for the DOS and DDOS attack which happens in the network [7]. The information gain and the gain ratio are the one which helps to find the top fifty features from the data set which helps to predict the attack.

The author has used the JRip classifier for classifying the data [8]. The most of the data that is stored in cloud requires the intrusion detection mechanism. It is very consuming process because the data set is high dimensional one where the best selection method is necessary for detecting the intrusions. The wrapper-based methods are selected for detecting the intrusions [10].Due to many redundant features in feature selection techniques it is very difficult to classify the intrusion in all aspects. Detecting all types of attacks by the single system is very costly. Three different data sets are used for intrusion detection. Hybrid feature selection method is used to classify the intrusions [11].

The feature selection process is the very important step in the intrusion detection. There are many threats that occurs in our daily life. The author has used the filter-based technique in research to increase accuracy and to reduce the time [12].

The RFF feature selection method is the one which uses the neural network to extract the feature information and trains the classifier to detect the intrusions. It is very useful in securing the system [13].

Methodology

The features like duration, protocol type, service, flag, urgent, host, count, serror_rate etc are the some of the samples that takes their roles in feature selection method.

The data set which is used to categorize the packet into normal or malicious. The IDS is the one which should be a generalized one and it should be able to categorize the attacks even for

the real time data set [14]. The feature selection method is the very important method for improving the model performance and to generate the good result. It is also the process which helps to reduce the problem of overfitting. It also intends to find the interpretability for the problem. There are various techniques for feature selection process [15].

Filter methods

The correlation-based method is one of the popular methods for feature selection in filter methods. It helps to identify the correlation between the two variables which gives more information about the target variable. It helps to reduce the redundancy in the dataset. The correlation matrix is the one of the statistical methods which helps to find relationship between two variables. It is used in many fields. Usually the correlation data will be taken in the form of table that is rows and columns. First, we need to find the correlation between the data in each cell of the table. It helps to summarize the large data with quick and easy way. The correlation value ranges from -1 to + 1. The -1 is the negative correlation and the + 1 is the positive correlation. If the value is 0 there is no correlation between variables.

It also predicts the independent variables how it is related to each other. The below is the formula for finding the correlation between the variables.

$$r = (n\Sigma XY - \Sigma X\Sigma Y) / \text{sqrt}((n\Sigma X^2 - (\Sigma X)^2)(n\Sigma Y^2 - (\Sigma Y)^2))$$

After computing the result, the value will range from + 1 to – 1. It will help to find the relationship and to decide the forecast for the data set. It will help to take good decisions. It is very easy to understand and read. In this highly correlated feature are removed. It will help to reduce the redundancy.

Statistical tests

It helps to find relationship between features and target variable.

Wrapper methods

It goes with selecting subset of features and there by analysing the performance of the model. This goes on by selecting all the subset of datathen selecting best subset which increases model performance.

Forward Selection: In this method each feature will be selected based on model performance.

Backward Elimination: It will start execution by initial features and later on least important feature will be eliminated.

Recursive feature Elimination: It willstart with initial set of features and then rank the features and repetitively eliminates the features until it meets it desired output.

Embedded Methods

The embedded methods will embedthe selection procedure in training the model. These machine learning model will only select the features.

LASSO Methods

It predicts the output based on the linear combination of features.

Tree Based Methods

It carries out the research based on providing the scores to features.

Dimensionality reduction methods

Dimensionality reduction is the process of reducing the features in the data set that is it will give the lower dimension of data.

Principal Component Analysis

This algorithm is one of the supervised learning algorithms which helps to reduce the dimension of the data set. When the dimension of data is very high then it will be very much difficult to obtain the results of classification and clustering. PCA will help to reduce the dimension. The dimensionality reduction will also help to reduce the input features without affecting the original information. The PCA will perform orthogonal transformation. The orthogonal transformation is nothing but it will convert the correlated variables to uncorrelated variables. It also helps to examine the interrelation between the set of variables.

t-Distributed Stochastic Neighbour Embedding

This method is also one of the dimensionality reduction problem which is used for visualization. It will help to reduce the high dimension data view to lower dimension view.

Information gain and Mutual Information

The information gain will help the user to identify the feature that contributes to the accurate prediction. The feature with highest information gain is choosed first in feature selection process. In decisiontree the feature with highest information gain is always choosed to reduce the dimension.

Cross Validation

The cross-validation technique will help to select the subset features in the data set which in turn will help to increase the efficiency of model performance.

Results and Discussion

In this approach PCA and the linear discriminant analysis are the two methods which are used for the dimension reduction and the feature selection process. The main objective of the PCA is to find the data with maximum variance. It is mainly used to reduce the complexity of the model. It is also used in pre-processing and visualization. The linear discriminant analysis helps in separation of the data between the classes. It will help to reduce the variance inside the classes. Features are selection based on their separation of classes. So, the LDA will focus on separation of data based on class labels. The PCA will focus on separation of variance. Both technics are very useful in separation of classes.

PCA is mainly choosed to find the variance in the dataset if IDS deviate from the normal patterns. The hybrid version of PCA LDA algorithm is used for dimensionality reduction. Initially PCA is used and later on LDA is used for dimensionality reduction.

Hybrid PCA LDA Algorithm

Step -1

$X \in \mathbb{R}^{m \times n}$ this is the original feature matrix where m is the number of instances and n is the number of features initially data standardization is done.

Step – 2

PCA for dimensionality reduction

Covariance matrix

$$S = \frac{1}{m-1} X_{\text{std}}^T X_{\text{std}}$$

Eigen value decomposition

Decompose S into eight vectors V and eigen values Λ

Selecting principal components

$$V_k = \begin{bmatrix} v_1 & v_2 & \dots & v_k \end{bmatrix}$$

Transforming data

Projecting the data on to principal components

Step -3

LDA for discriminative feature selection

Label the data:

Label the data either it as normal or intrusive

Compute Class Means

Calculate the mean vectors for each class

$$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{\text{pca}_j}^{\text{class } i}$$

Within class scatter matrix SW

Calculate the within class scatter matrix

$$S_W = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{\text{pca}_j}^{\text{class } i} - m_i)(X_{\text{pca}_j}^{\text{class } i} - m_i)^T$$

Between class scatter Matrix

Calculate the between between class matrix

$$S_B = \sum_{i=1}^{c} n_i (\bar{m} - m_i)(\bar{m} - m_i)^T$$

Generalized Eigen value problem

Solve the eigen value problem to find the eigen values

$$S_W^{-1} S_B w = \lambda w$$

Selecting Discriminant Features

Choose the top k eigen values corresponding to k largest eigen values

Transforming Data

Project the PCA transformed data onto the selected discriminant features

Step 4

Model Training and Evaluation

Train the Classifier

The reduced dimensional data can be used to train the classifier

Evaluation Metrics

The evaluation of metrices is done to check the classifier performance through cross validation.

Step – 5

Fine Tuning and Validation

Based on the performance of the model fine tune the principal components and the discriminant features.

Validation

Validate the model to ensure the generalizability

Deploy the model

Deploy the model for the real-world data set.

Experimental Results

The correlation matrix for the IDS data set is shown.

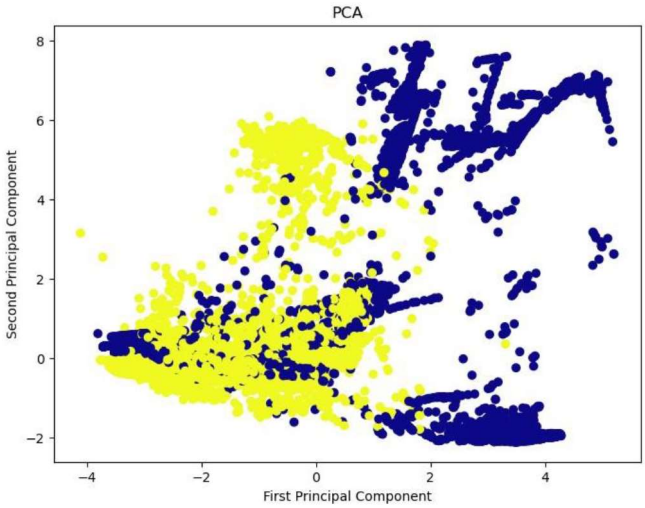

Fig 1: Correlation Matrix

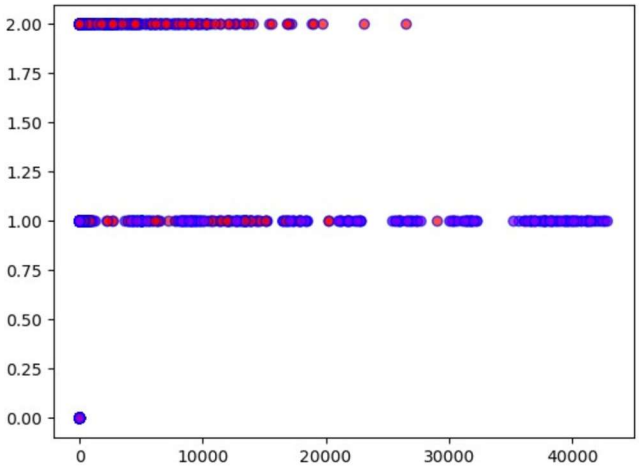

Fig 2: LDA Visualization of the IDS Data set

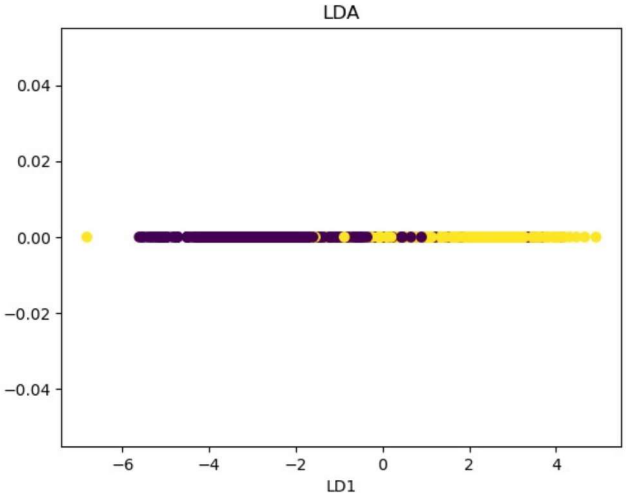

Fig 3: LDA Visualization of the IDS Data set

The algorithm is compared with other machine learning models to validate the performance. The method is compared with various other method like ada boost, bagging, logit boost, voting, stacking and iterative classifier. It is used to increase the performance of the classifier.

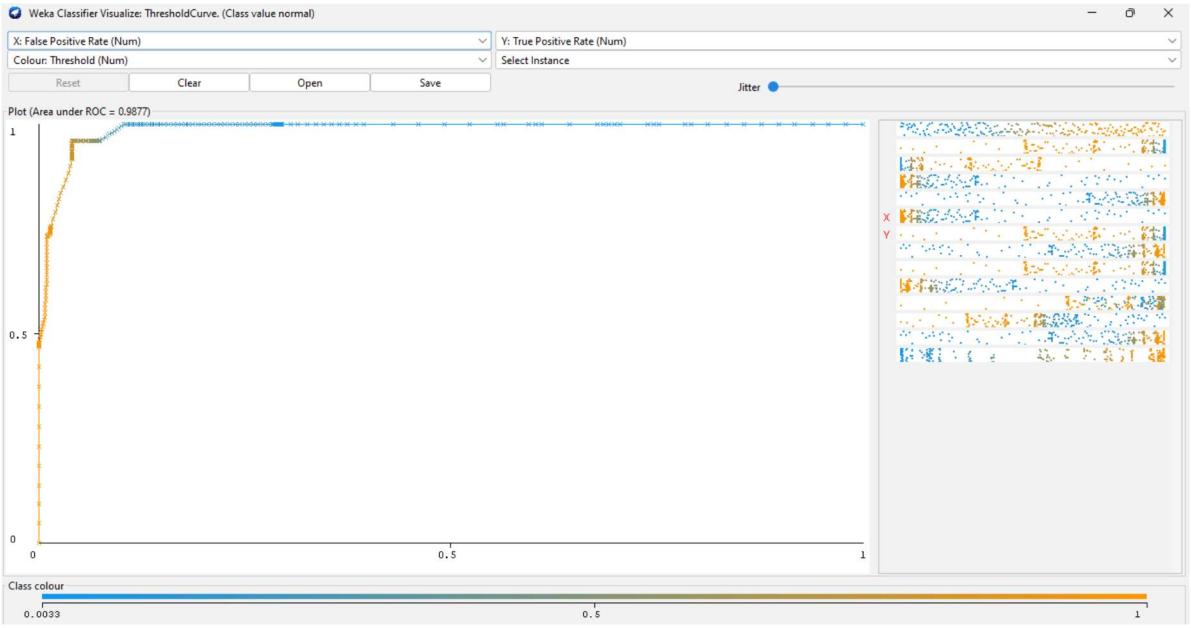

Fig :4 ROC curve of Ada Boost with PCA-LDA Feature selection

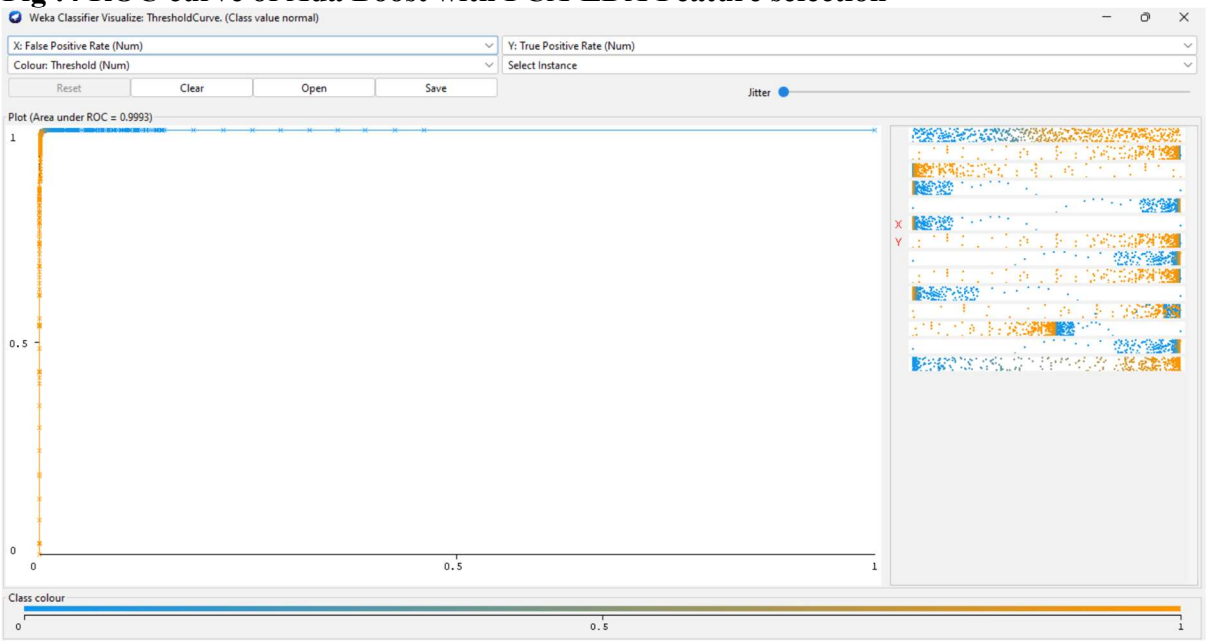

Fig :5 ROC curve of Bagging with PCA-LDA Feature selection

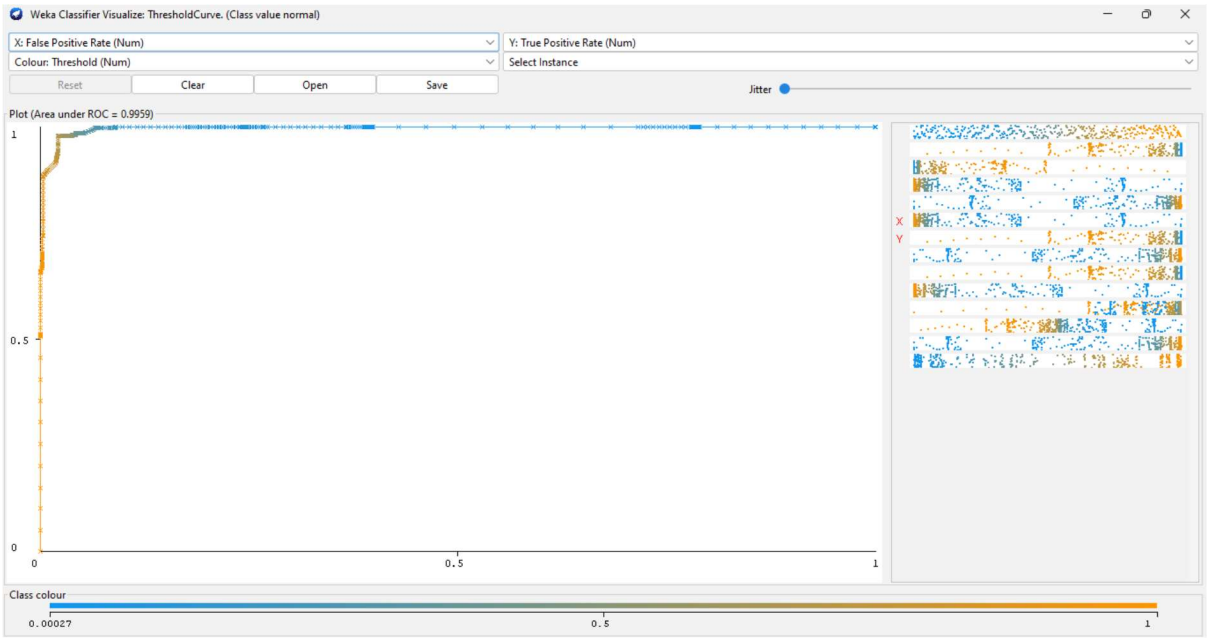

Fig :6 ROC curve of Logit boost with PCA-LDA Feature selection

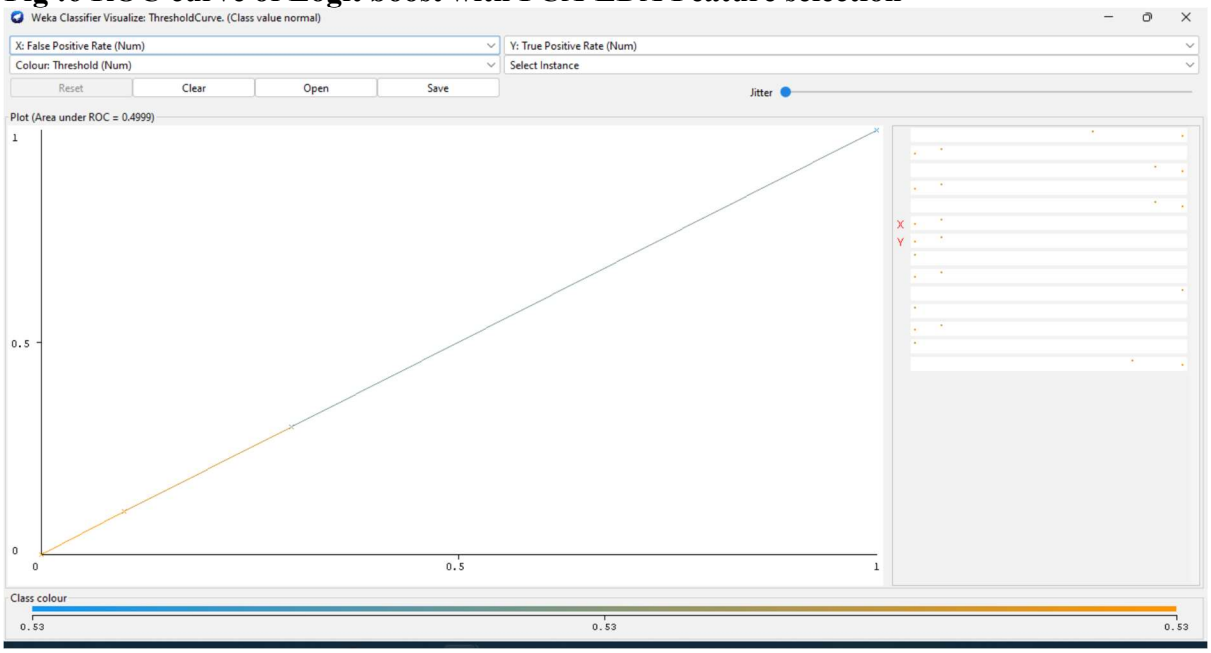

Fig :7 ROC curve of Voting with PCA-LDA Feature selection

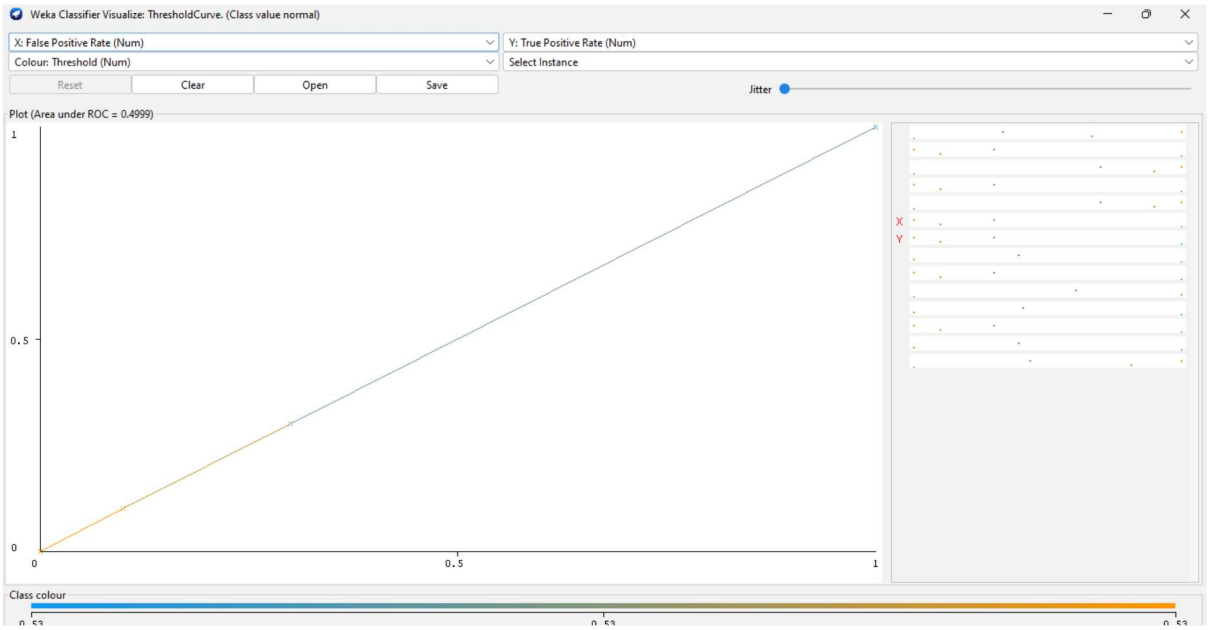

Fig :8 ROC curve of Stacking with PCA-LDA Feature selection

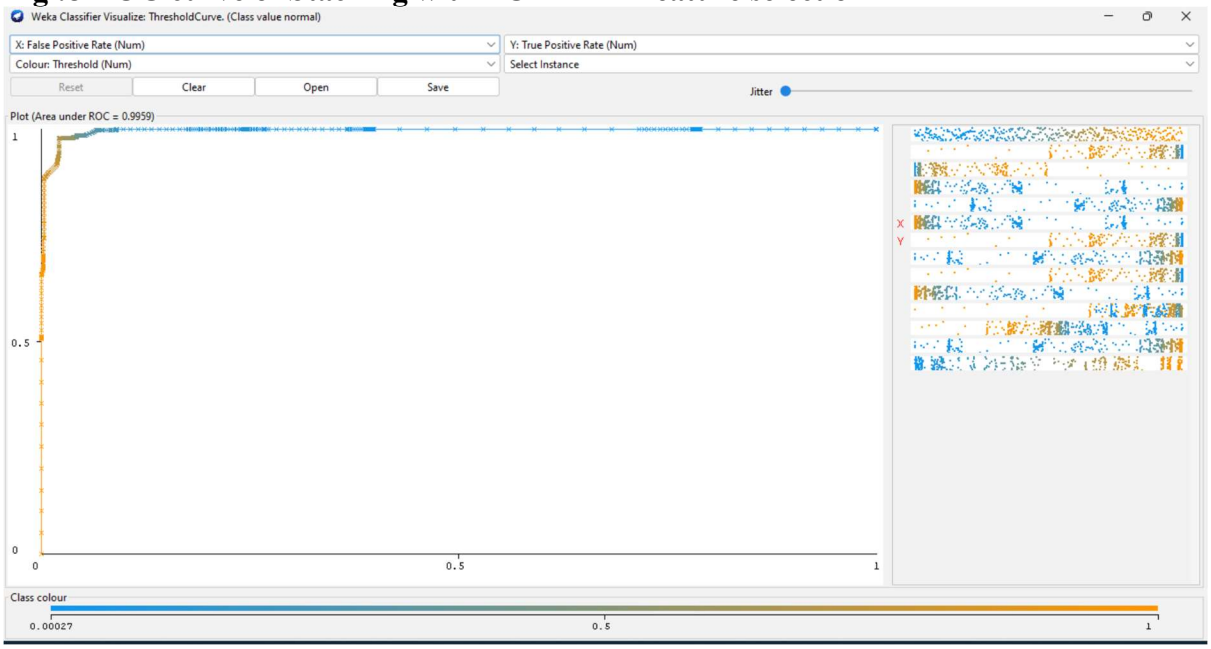

Fig :9 ROC curve of Iterative Classifier Optimizer with PCA-LDA Feature selection

Finally,while comparing the machine learning models Adaboost, Bagging, Logit Boost, Iterative Classifier shown the best results in classifying the normal and anomaly instances.

Conclusion

There are various feature selection methods which helps to choose the important features and there by it helps to increase the accuracy and the time. The PCALDA hybrid algorithm helps to perform the good results while comparing with other machine learning classifiers. The classifiers like Adaboost, Bagging, Logit Boost, and Iterative classifier optimizer performs well. It effectively classifies the instances either it as normal or anomaly.

References

[1] Dandy PramanaHostiadi; YohanesPriyoAtmojo; Roy Rudolf Huizen; I Made Darma Susila; GedeAnggaPradipta; I Made Liandana, "A New Approach Feature Selection for Intrusion Detection System Using Correlation Analysis", IEEE Explore, 2022.[2] Hussein Al Zoubi ,SamahAltaamneh , " A feature selection technique for network intrusion detection based on the chaotic crow search algorithm", IEEE Explore, 2022.[3] Mohamed Hammad, Wael Medany , Yasser Ismail, " Intrusion Detection System using Feature Selection with Clustering and Classification Machine Learning Algorithms on the UNSW- NB 15 dataset " , IEEE Explore, 2020.[4] Arar Al Tawil, KhairEddin Sabri, " A feature selection algorithm for intrusion detection system based on Moth Flame Optimization", IEEE Explore, 2021.[5] A Lakshmanarao A Srisaila , T Srinivasa Ravi Kiran," Machine Learning and Deep Learning framework with feature selection for intrusion detection", 2022, IEEE Explore.[6] Saikat Das, SajalSaha, Annita Tahsin Priyoti, EteeKawna Roy, Frederick T Sheldon , Anwar Haque and Sajjan Shiva , " Network Intrusion Detection and Comparative Analysis using Ensemble Machine Learning and Feature Selection", IEEE, 2022.[7] Muhammad HilmiKamarudin, Carsten Maple and Tim Watson, "Hybrid Feature Selection technique for intrusion detection system", International Journal of High-Performance Computing and Networking, Vol 13, No 2, 2019.[8] V.R Balasaraswathi, L Mary Shamala ,Yasir Hamid, M Pachhaiammal Alias Priya, M Shobana, MuthukumarasamySugumaran, " An efficient feature selection for Intrusion Detection System Using BHKNN and C2 Search Based Learning Model ", ACM Digital Library, Volume 54, Issue 6, 2022.[9] PushparajNimbalkar, Deepak Kshirsagar, "Feature Selection for Intrusion detection System in Internet of Things (IOT) ", Science Direct,Volume 7, Issue 2,2021. [10] Sanjay Razdan, Himanshu Gupta, Ashish Seth, "Feature Selection Methods for Intrusion Detection Systems : A Performance Comparison" , IEEE Explore, 2022.

[11] Ebrime Jaw, Xueming Wang, " Feature Selection and Ensemble Based Intrusion Detection System : An Efficient and Comprehensive Approach", MDPI, 2021.

[12] Yang Lyu, Yaokai Feng, and Kouichi Sakurai, " A Survey on Feature Selection Techniques Based on Filtering Methods for Cyber Attack Detection ", MDPI, 2023.

[13] Kezhou Ren, Yifan Zeng Zhiqin Cao &YingchaoZhang ," ID – RDRL : A deep Reinforcement learning based feature selection intrusion detectionmodel " , Scientific Reports, 2022.

[14]Hadeel Alazzam, Ahmad Sharieh, KhairEddinSabri, "A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer ", Elsevier, 2020.

[15]MerveOzkan Okay , RefikSamet, Omer Aslan SelahattinKosunalp , TeodorLliev, and IvayloStoyanov, " A Novel Feature Selection Approach to Classify Intrusion Attacks in Network Communications " Applied Sciences, 2023.