Linear Discriminant Analysis based Hybrid SVM-CART for Intrusion Detection System

[1]Navdeep, [2]Ishpreet Singh Virk
Baba Banda Singh Bahadur Engineering College
Fatehgarh Sahib,Punjab

Abstract **-Intrusion Detection System has been a major problem of research for a long time. Researchers have tried different clustering algorithms for accurately classifying the intrusion detection system data so that it can be utilized in the real time monitoring. In this research work a Linear Discriminant Analysis based Support Vector Machine-CART algorithm for classification intrusion detection system data is implemented. The linear discriminant analysis is used for accurately predicting the important features from the highly dimensional data. Reducing the dimensions of the data improves the computation time. A linear line is drawn as vector and the classification is done giving more weightage to the SVM for data points which are close to the line. For distant points, the CART algorithm is given more weightage. The hybrid algorithm performs quite better than the other algorithms and it is found to give better results in terms of accuracy, precision, recall.**

1. Introduction

Data Security is area of concern since internet has discovered. The Small hole (error) to database can result to the stop the growth of company. Data needs to secure offline as well as online. Offline protection of the data can be done by passwords etc. For Protection of the data against internet attacks, intrusion detection is mandatory.IDS are of two types one is which detect new type of attack while other may detect the attack from the known set of patterns. Intrusion detection actually aims to improve the rate of the detection of attacks.

The basic work of the intrusion detection is collection of the data i.e. the various activities carried out on our machine by the other machines and then filtering the malicious activities. The attack pattern or information is recorded for the security purposes, so that it can be used in future for detection of intrusion. The recorded attack pattern helps in establishing the proper intrusion prevention system for the attack. It traces user activity from the point of entrance to point of impact. As the attack is detected, its pattern is recorded and stored as a signature in the Intrusion Detection System. Thus, overtime, IDS improves continuously and concurrently, IDS is able to detect more attacks. On the other hand, the attacker may be aware of the recorded patterns. Hence, the attacker would use the new way or new pattern to attack the system. Therefore, there is no ideal intrusion detection system. One more important lagging behind point of IDS is that it cannot operate on its own. When installed on a system, an administrator is required to monitor the IDS.

Hence, IDS is an important part of the data and network security. So, to create ideal IDS for each type of the problem domain, various experiments are to be considered. Researchers have done study for five decades over this topic. Continuous research and the study of the topic over the years have improved the results. But, to consider any intrusion detection system as a perfect one is not possible.

There are many classification and clustering techniques are used in intrusion detection system.Classification techniques are Linear Classifiers,Support Vector machines,Quadratic tress,decision trees,neural network etc. Clustering aims to combine the data and the clusters are formed on the basis of similarity and dissimilarity of the data objects.The Clustering can be done using the distance as the similarity measure, whereas some use density based clustering. Clustering techniques are K-means clustering, Fuzzy c-means clustering, Hierarchical clustering etc.We use Hybrid SVM-CART Classification technique with Linear Discriminant Analysis.

Load Discriminant Analysis is applied on the dataset and a weight factor is calculated which is a m x n matrix where m is the number of classes and n is the number of attributes. The weight matrix is multiplied with each class of the data and the reduced data is saved into a new file.

Linear Discriminant Analysis is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-reparability in order avoid over fitting and also reduce computational costs.

LDA is closely related to analysis of variance and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. Linear discriminant analysis has continuous independent variables and a categorical dependent variable.

The Support Vector Machine has attracted a high degree of significance in the machine knowledge research community . Several recent studies have reported that the SVM (support vector machines) normally are capable of delivering higher presentation in terms of classification accuracy than the previous data classification algorithms. Sims have been in employment in a wide range of real world problems such as text organization, hand-written digit recognition, tone recognition, image categorization and object detection, micro-array gene expression data analysis, data classification. It has been shown that Sims is consistently superior to other supervised learning methods .

Classification and Regression trees were introduced by Breiman et al in 1984. The major idea behind tree technique is to recursively partition the data into smaller and smaller strata in order to get better the fit as best as possible. They partition the sample space into a set of rectangles and fit a model in each one. The sample space is originally split into two regions. The optimal split is found over all variables at all possible split points. For every of the two regions created this process is repeated again. Hence some researchers have termed the technique recursive partitioning. The main components of the CART methodology are the selection and stopping rules. The selection rule determines which stratification to perform at every phase and the stopping rule determines the final strata that are formed. Once the strata have been created the impurity of each stratum is measured. The heterogeneity of the outcome categories within a stratum is referred to as "node impurity". Classification trees are employed when the outcome is categorical and regression trees are employed when the outcome is continuous. Classification trees can take most forms of categorical variables including indicator, ordinal and non-ordinal variables and are not limited to the analysis of categorical outcomes with two types.

Decision tree construction is a well-known method for classification. A file for decision tree categorization consists of a set of data files, which are pre-classified into q(≥ 2) known module. The reason of conclusion tree construction is to partition the data to separate the q classes. A conclusion tree has two types of nodes, decision nodes and leaf nodes.

2. Review of Literature

Srivastava and Bhambu [7] defined that classification is one of the mainly important tasks for different function such as text classification, tone recognition, image classification, micro-array gene expression, proteins structure predictions, data categorization etc. Mostly of the existing supervised classification technique are based on established statistics, which can give ideal results when sample size is treatment to infinity. However, only finite samples can be acquired in way. In this paper, a learning technique, Support Vector Machine (SVM), is useful on dissimilar data (Diabetes data, Heart Data, Satellite Data and Shuttle data) which have two or multi class. SVM, a powerful machine way industrial from statistical learning and has complete important achievement in some field. Introduced in the early 90's, they led to an explosion of interest in machine learning. The foundations of SVM is developed by Vapnik and are attainment popularity in field of machine learning due to a mixture of attractive kind and promising empirical show.

Bettencourt and Clarke [8] describes that hyper-spectral isolated sensing increases the volume of in order obtainable for research and research, but brings with it the need for resourceful numerical technique in sample spaces of many dimensions. Due to the complexity of problems in elevated dimensionality, some ways for dimension decrease are optional in the literature, such as major Components Analysis. Although PCA can be functional to data decrease, its use for classifying images has not created good results. In the current study, the Classification and Regression Trees process, more widely known by the acronym CART, is used for quality selection. CART involves the identification and construction of a binary decision tree using a sample of training records for which the accurate classification is known. Binary decision trees consist of repeated divisions of a quality space into two sub-spaces, with the terminal nodes related with the classes. A desirable decision tree is one having a relatively small number of branches, a relatively small number of intermediate nodes from which these branches diverge, and high analytical power, in which entity are accurately classified at the terminal nodes.

Gordon [10] states thatan organization and regression tree is a non-parametric methodology first introduce by Breiman with colleagues in 1984. The employed using two are SAS Enterprise Miner and several for examples are given to demonstrate their use in this work. They are underused and they have the facility to divide populations into significant subgroups which will allow the arrangement of groups of interest and recover the of products and services accordingly. They provide the simple powerful examination.

Kaur and Kaur [4] defined that the overall objective of the data mining way is to extract data from a large data set and transform it into an understandable form for further use. Clustering is importantfor data analysis and data mining applications. It is task of grouping set of the objects so that objects in the same group are more same to each other groups. There are dissimilar types of clusters: Well-separated clusters, Center-based clusters, Contiguous clusters, Density-based clusters, Shared Property or Conceptual Clusters. Predictive and the descriptive are the two major tasks of the data mining. Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering; the value of k-mean is set. Density based clusters are definite as area of higher density then the remaining of the data set. Grid based clustering is processing time that normally depends on the size of the grid instead of the records. The grid based methods use the single uniform grid mesh to mainly problem domain into cells. In this survey paper, a review of clustering and its different methods in data mining is done.

Kanungo and Mount [11] statesthat k-means clustering is set of n data points in d-dimensional space Rd and an integer k. The main problem is a set of k points in Rd, called centers, so they reduce the mean squared distance from every data point to its nearest center. An accepted heuristic for k-means clustering is Lloyd's algorithm. In this paper, they used the simple and efficient implementation of Lloyd's k-means clustering algorithm, they are called the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only chief data structure. The practical efficiency is creating the filtering algorithm in two methods. First methods, they present a data-sensitive analysis of algorithm's running time, which shows that the algorithm runs in advance as the separation between clusters increases. Second, they studies both on synthetically created data and on real data sets from applications in colour quantization, data compression, and image segmentation.

Anand and Patel [3] defined that Intrusion Detection System is used the vital instrument in the defending network from the malicious or abnormal activity. It is still popular is know, what intrusions have happened or are happening, so we can understand the security issues or risks and we enhanced prepared for future attacks. The ability to analyze network traffic, recognize incoming and ongoing network attack, common network administrator has turn to IDS to help are detecting anomalies in network

traffic. In this paper, we study dissimilar types of attacks on IDS and gives a description of dissimilar attack on different protocol like TCP,UDP,ARP and ICMP.

Yingzho and Karypis [8] states that quick and high-class document clustering algorithms play an significant task in supply intuitive navigation and other browsing mechanisms by organizing large amounts of data into a low number of meaningful clusters. The clustering algorithms that build key hierarchies out of very large document collection are ideal tools for their interactive visualization and study as they give data-views that consistent, predictable, and at the dissimilar levels of granularity. This paper focus on the document clustering algorithms and that build such hierarchical solutions. and first are the presents a comprehensive study of partitioned and agglomerative algorithms that utilize dissimilar criterion functions and merging schemes, and next are the present a new class of clustering algorithms is also called constrained agglomerative algorithms,.

Burges [6] suggest that the study starts an overview of the idea of VC dimension and structural risk minimization. They explain linear Support Vector Machines for separable and non-separable data, working through a non-trivial for example in feature. They explain a mechanical analogy, and consider when SVM solutions are single and when they are global. They show how keep vector training can be practically implemented, and study details the kernel mapping way. Which is used to construct SVM solutions are nonlinear data? And show how to Support Vector machines can have very large VC dimension by computing the VC dimension for homogeneous polynomial and Gaussian radial are basis function kernels.

Yasinsac and Goregaoker [1] states that the Internet has emerged as middle for wide-scale electronic communication connecting financial transactions and other sensitive information. Encrypted interactions between principal are normally used to ensure data security. Security protocols are rules that govern such encrypted exchanges. This paper describes a system for detecting intrusions on encrypted exchanges over public networks by identify the type of security protocols and attacks on them.

Fawcett [12] states that Receiver operating characteristics (ROC) graphs are useful for organizing classifiers and visualizing their performance. Roc graphs are able to provide a richer measure of classification performance than scalar measures such as accuracy,error rate or error cost.Because they decouple classifier performance from class skew and error costs,they have advantages over other evaluation measures such as precision-recall graphs and light curves.

Qiao et al [13] describe Fisher's linear discrimental analysis is typically used as a feature extraction or dimension reduction step before classification. It finds the projection directions such that for the projected data, the between class variance is maximized relative to the within-class variance. Once the projection directions are identified, the data can be projected to these directions to obtain the reduced data, which are usually called discriminant variables. These discriminant variables can be used as inputs to any classification method, such as nearest centroid, k- nearest neighborhood and support vector machines.

Santra and Christy [2] states that confusion matrix is more commonly used named contingency table in which the matrix could be arbitrarily large, the number of correctly instances is the sum of diagonals in the matrix,all others are incorrectly classified accurately.

Janardan et al [8] states Linear Discriminant Analysis is a well-known scheme for feature extraction and dimension reduction. It has been used widely in many applications involving high-dimensional data, such as face recognition and image retrieval. An intrinsic limitation of classical LDA is the so-called singularity problem, that is, it fails when all scatter matrices are singular. A well-known approach to deal with the singularity problem is to apply an intermediate dimension reduction stage using Principal Component Analysis (PCA) before LDA. The algorithm, called PCA+LDA, is used widely in face recognition. However, PCA+LDA have high costs in time and space, due to the need for an eigen-decomposition involving the scatter matrices.

Welling states that this technique searches for directions in the data that have largest variance and subsequently project the data onto it. In this way, they obtain a lower dimensional representation of the data, that removes some of the "noisy" directions. There are many difficult issues with how many directions one needs to choose, but that is beyond the scope of this note.

Vinchurkar and Reshamwala [4] reviewed intrusion detection techniques square measure. In this paper they need basic structure of intrusion detection system and conjointly describe best intrusion detection system. They need intrusion detection system on premise of the information supply and model of intrusion. Varied challenges in intrusion detection square measure well given. Neural network machine learning approach for intrusion detection square measure evaluated. They conjointly create dimension reduction victimization PCA. They reach conclusion this IDS can be effective to update the audit knowledge quick and conjointly, there's have to compelled to style IDS that may be with challenges of huge info and rising performance measures.

3. Research Methodology

There has been a lot of work in the field of intrusion detection system. The research work intends to develop a hybrid algorithm of Linear Discriminant Analysis based Support Vector Machine-Classification and Regression Tree. The SVM classifier uses a support vector along the main component and classifies various data into different clusters based on the values of support vector. The Algorithm will be hybrid with CART algorithm which is based on the regression tree concept. CART is classification method which uses historical data to construct decision trees. Depending on available information about the dataset, classification tree or regression tree can be constructed. Constructed tree can be then used as dataset for classification of new observations. Classification trees are used when for each observation of learning sample we know the class in advance. Classes in learning sample may be provided by user or calculated in accordance with some exogenous rule. For example, for stocks trading project, the class can be computed as a subject to real change of asset price. The LDA is used for feature extraction from the large number of features, which will reduce the computation cost. In this research work an idea of intrusion detection system using SVM-CART aided by Linear Discriminant Analysis for dimensionality reduction is implemented.

Steps

1. Data set is obtained from the KDDCUP99.
2. Linear Discriminant Analysis is applied for reduction of dimensionality of the data and selects the best features out of it.
3. Classification algorithm is applied on the data.

4. Support Vector Machine is given more weightage for nearby points to the line and Regression tree (CART) is given more weightage for distant points
5. Activities will be monitored and parameters will be calculated

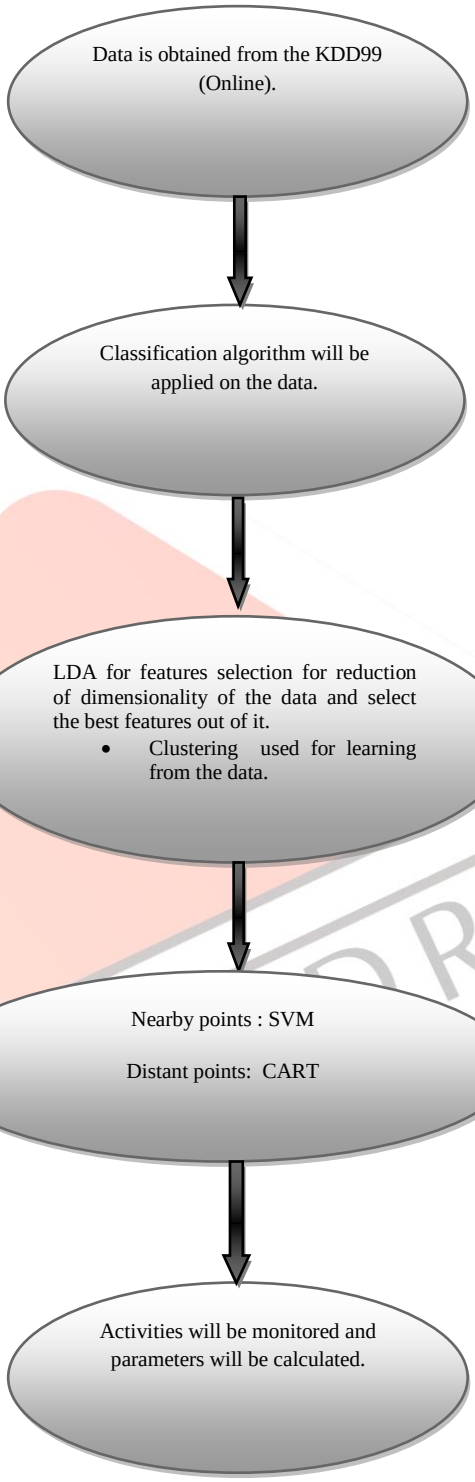
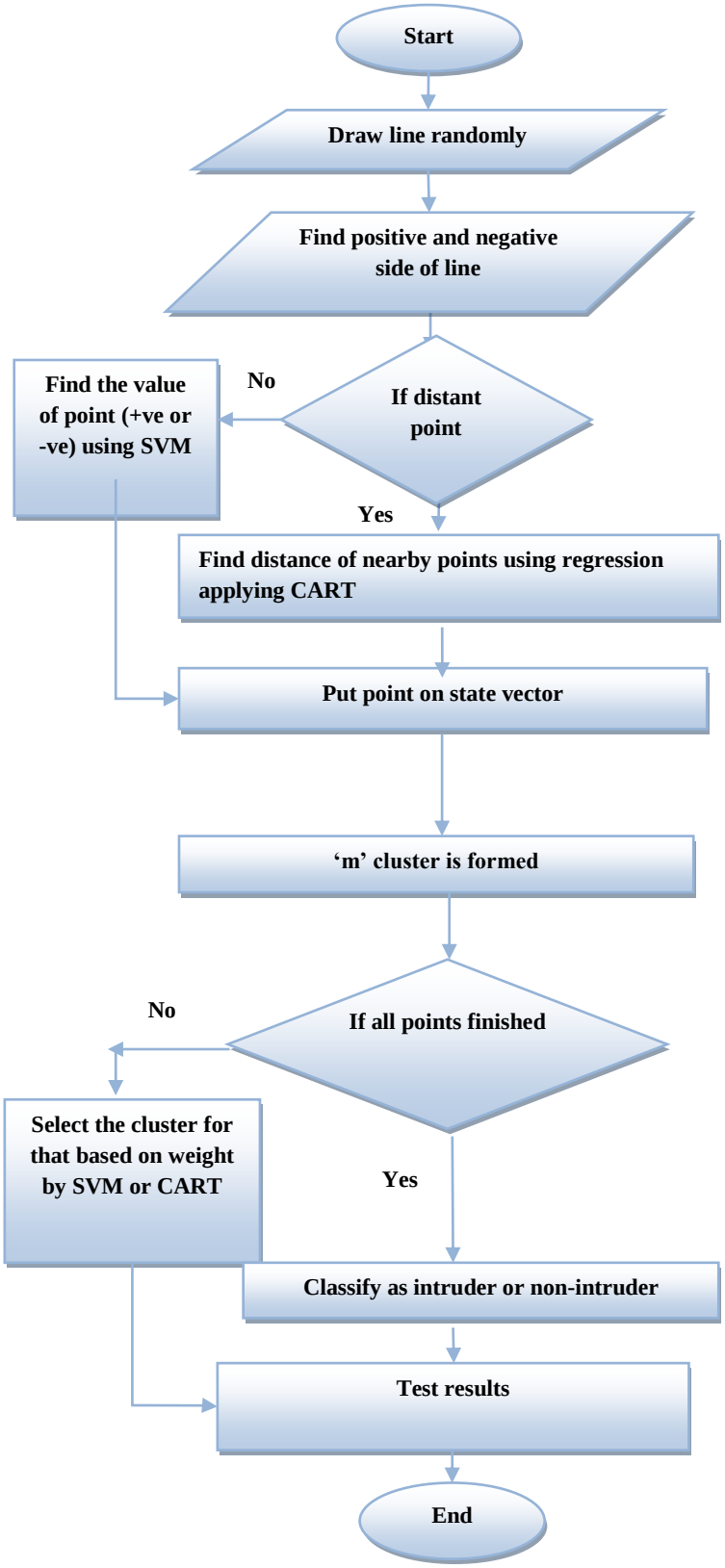


Figure 3.1: Method of Work.

3.1 Flow chart of LDA based SVM-CARTAlgorithm:


Start
Draw line randomly
Find positive and negative side of line
If distant point
No
Find the value of point (+ve or -ve) using SVM
Yes
Find distance of nearby points using regression applying CART
Put point on state vector
'm' cluster is formed
If all points finished
No
Select the cluster for that based on weight by SVM or CART
Yes
Classify as intruder or non-intruder
Test results
End

4.Results and Discussions

To examine the result there are two types of method one is confusion Matrix and other is ROC analysis.

4.1 Confusion Matrix

Confusion Matrix is more commonly used named contingency table in which the matrix could be arbitrarily large, the number of correctly instances is the sum of diagonals in the matrix, all other are incorrectly classified accurately .

TP	FN
FP	TN

4.2 ROC Analysis

Receiver operating characteristics (ROC) graphs are useful for organizing classifiers and visualizing their performance. Roc graphs are able to provide a richer measure of classification performance than scalar measures such as accuracy,error rate or error cost.Because they decouple classifier performance from class skew and error costs,they have advantages over other evaluation measures such as precision-recall graphs and light curves.

To examine the methodology, stated in this paper, the whole algorithm and simulation is done in the python environment using standard KDDCUP 99 dataset generally used for intrusion detection system. This is discovered by the Stlofo et al[ ]. and is based in the perspective of the information caught in DARPA 98' intrusion detection system. Training dataset is being used at a sampling rate of 100 of the actual training set and testing set is used at the rate of 64 of actual testing set.

4.3Evaluation Parameters

The evaluation parameter to measure the efficiency of algorithm is Accuracy, Detection Rate and False Positive Rate which is governed by the confusion matrix. Confusion matrix is calculated at the end of the simulation. The confusion matrix contains True Positive (TP), True Negative(TN), False Positive(FP) and False Negative(FN) of the every class. The mentioned parameters are calculated using the above mention terminologies which is discuss here.

4.3.1 Accuracy:

It represents the total no of accurately predicted samples w.r.t. the total no of samples tested. More is the accuracy better is the method. Accuracy can be computed as:

$$Accuracy = TP/(TP + FN) \dots\dots\dots\dots\dots\dots\dots\dots (1)$$

4.3.2 Detection Rate:

It is defined as the total no. of samples predicted of particular class out of the total samples of that class i.e.

$$Detection\ Rate = \frac{TP}{TP + FP} \dots\dots\dots\dots\dots\dots\dots\dots (2)$$

4.3.3 False Positive Rate:

False Postive Rate provides number of the normal queries that can be detected by mistake as queries belongs to particular class:

$$False\ Positive\ Rate = \frac{FP}{FP + TN} \dots\dots\dots\dots\dots\dots\dots\dots (3)$$

4.3.4 Sensitivity

$$Senstivity = Recall = \frac{\mathrm{TP}}{\mathrm{P}} \dots\dots\dots\dots\dots\dots\dots\dots\dots (4)$$

4.3.5 Specificity

$$Specificity = \frac{TN}{FP} + TN \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (5)$$

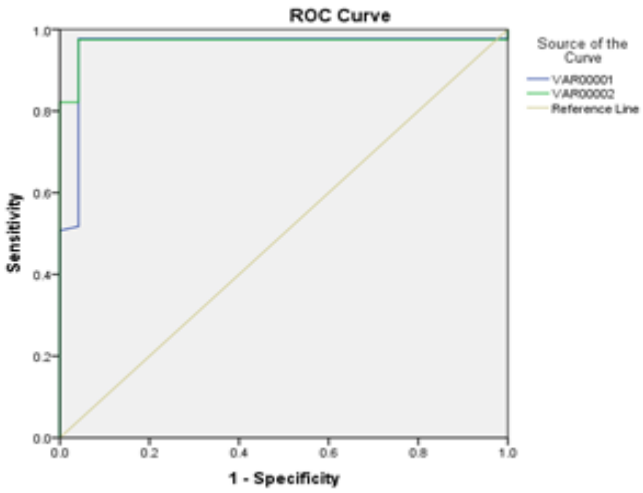
4.3 Confusion Matrix and result of evaluation parameters for each class is shown in table 1 and 2 respectively

Table 4.1

Actual class	Predicted Class		
	Normal	DoS	Probe
Normal	905	21	4
DoS	95	3497	0
Probe	10	15	32

Table 4.2

Parameters	Classes		
	Normal	DoS	Probe
Accuracy	0.97	0.97	0.56
Detection rate	0.89	0.98	0.88
False positive rate	0.28	0.03	0.008



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
VAR00001	.958	.020	.000	.918	.998
VAR00002	.968	.010	.000	.949	.988

The test result variable(s): VAR00001 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

6.Conclusion

Implementation of the methodology of SVM-CART classification on the standard KDDCUP 99 dataset has shown good result. As we can see from the above figures that the false positive rate of the not exceeding 0.1 for all the classes which shows efficacy of the algorithm. Also the detection rate of all the classes is more than the 75 % of the provided test samples. The result also shows the better accuracy for most of the classification. The future scope for the methodology is to use the corrected dataset, improving the results for normal queries and also the model can be tested for the parameters other than the listed in the paper.

7.Reference

[1] Alec Yasinsac and SachinGoregaoker, "An Intrusion Detection System for Security Protocol Traffic".
[2] A.kSantra and C.JosephineChristy,"Genetic Algorithm and confusion Matrix for Document Clustering",International Journal of Computer Science issues,January 2012,vol.9,pp. 322-328.
[3] Amrita Anand and Brajesh Patel, "An Overview on Intrusion Detection System and Types of Attacks It Can Detect Considering Different Protocols", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 8, August 2012.
[4] Amandeep Kaur and Navneet Kaur, "Survey Paper on Clustering Techniques", IJSETR, Volume 2, Issue 4, April 2013.
[5] AlpaReshamawala and Deepika P. Vinchurkar,November 2012 "A Review of Intrusion Detection System using Neural Network and Machine learning Technique" International journal of Engineering Science and Innovative Technology, Vol.1, Issue 2.
[6] Christopher J.C Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Kluwer Academic, 1998.
[7] Durgesh K. Srivastava and LekhaBhambhu, "Data classification using support vector Machine", Journal of Theoretical and Applied Information Technology, 2009.
[8] H. R. Bittencourt and R. T. Clarke, "Feature Selection by using classification and regression trees (CART)" , 2000.
[9] Jieping Ye, RaniJanardan and qili,"Two-Dimensional Linear Discrimenant Analysis".
[10] Leonard Gordon, "Using Classification and Regression Trees (CART) in SAS Enterprise Miner For Applications in Public Health", SAS Global Forum 2013.
[11] Tapas Kanungo and David M. Mount, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE, VOL. 24, NO. 7, JULY 2002.
[12] Tom Fawcett,"An Introduction to ROC analysis", Science Direct,December 2005.
[13] ZhihuaQuiao,Lan Zhou and JinnhuaZ.Huang,"Effective Linear Discriminant Analysis For High Dimensional Low Sample Size Data".