Intrusion Detection System Using SVM

[1] Asmita Bhaskar Tilave, [2] Madhavi Raghunath Trimbakkar, [3] Shyam Nitin Shinde
[123]UG Student(Computer Engineering Department),
[4]Prof.Swapnil Brijalal Kadam, [5]Prof. Nilam Nisar Shaikh,
[45]Assistant Professor(Computer Engineering Department),
SSPM's College of engineering, Kankavali, India. 416602

Abstract : In recent years, there are modest changes in growth of computer network likelihood of network security. The rapid increasing unauthorized activities in network, traditional firewall techniques cannot provide complete protection against attacks or intrusion. Intrusion Detection System (IDS) is one of the popular techniques for security. IDS detect the complicated intrusions in networks. In this paper, we present a Support Vector Machine (SVM) approach with feature scaling to detect various types of network intrusion. The experimental result shows higher detection accuracy rate and reduces the time complexity using kernel SVM.

IndexTerms **- intrusion detection system, support vector machine, KDD_dataset, Feature scaling**

I. INTRODUCTION

Now days, rapidly development is internet service and therefore, network security is viral and important issue in network environment. Intrusion detection system is one of best application for oppose the external attacks or intrusion. An intrusion detection system examines network traffic to find out the intrusion as it happen in the network and report them to the network administrator. In IDS vulnerability based on the network administrator because if network administrator is not present then IDS can't be able to generate the alert message.

There were so many techniques are developed for network security like firewall which is more powerful and useful technique. But you think about why we use IDS instead of firewall. So the main reason of using IDS is that firewall can only protect the network from outside attacks it cannot detect the intruders in the network whereas IDS can detect them. Also firewall cannot check every packet in the network whereas each and every packet is scan by IDS. The challenging issues of IDS are information security and detection of security threat.

The real time example of intrusion detection system is burglar alarm which is installed at our home. The use of lock system is to protect the house from theft. If someone wants to break the lock system and enter into the house then burglar alarm detect the lock has been broken and alert them by generating alarm.

II. RELATED WORK

Pavan Kaur and Dr. Dinesh Kumar [2015] have proposed a hybrid model for feature selection and intrusion detection is the most important issue in IDS. The selection of feature and attack attribute and normal traffic attribute is challenging task. The main objective of this paper is to detect from network from different dataset using KNN, GA and SVM in Weka tool. The selections of different features are based on KDD CUP'99 benchmark dataset. The performance of this work measured in detection rate, computational time and root mean square error.

Adetunmbi A.Olusola, Adeola S.Oladele and Daramola O.Abosede [2010] have proposed, Analysis of KDD'99 intrusion detection dataset for selection of relevance features to the detection of each class. For selecting important features from input data lead to a simplification of the problem and faster as well as more accurate detection rate. A rough setdegree of dependency and the dependency ratio of each class we reemployed to determine the most discriminating features for each class.

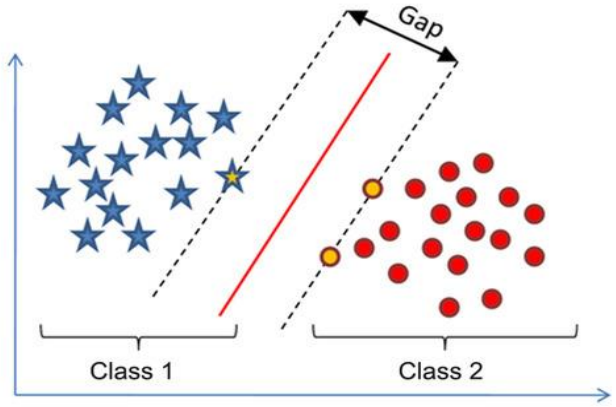
Dr. S.Vijayarani1 and Ms. Maria Sylviaa [2015] proposed, Intrusion Detection System- A *Study'* This main objective of this papers to provide a complete study about the definition of intrusion detection, history, life cycle, types of intrusion detection methods, types of attacks, techniques and different tools, research needs, applications and challenges.

Nilotpal Chakraborty published international journal in 2013 on 'Intrusion Detection System and Intrusion Prevention System: A Comparative Study' which proposes comparative study of intrusion detection system and intrusion prevention system their functionality, their performances and their effectiveness to malicious activity over the network.

III. SUPPORT VECTOR MACHINE

What is SVM?

Support vector machine (SVM) is the supervised learning model which performs both regression and classification of the tasks. But, it is mostly used for classification problems. SVM constructs a hyper plane or set of hyper planes in n-dimensional space, which can be used for classification or regression purpose. Hyper plane separate the features into two or more classes, as shown in fig 1.



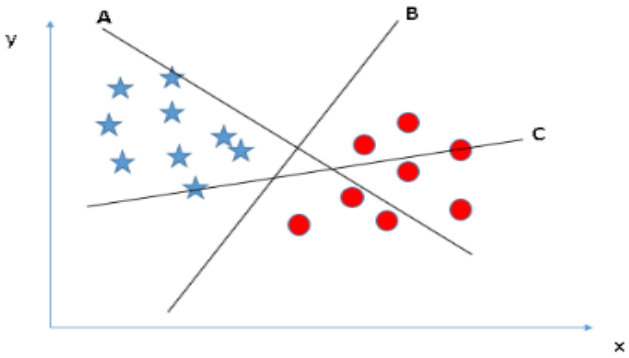
Support vector machine

Hyperplane gives a good separation that has the largest distance to the nearest training data point of any class.

How does it work?

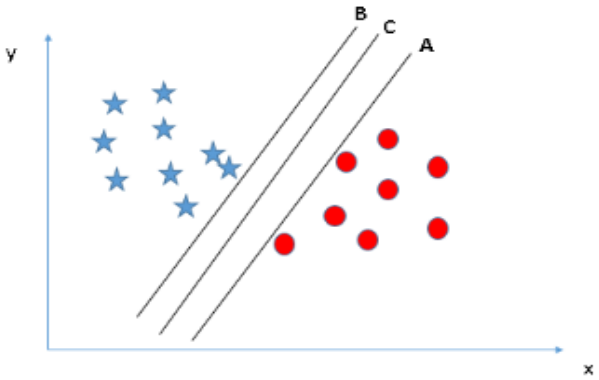
In above question, we got idea about hypherplane that separate the two classes. Right now, we have another question is "which hypherplane is right?"

Scenario-1: Here, we have three hyper planes (A, B and C). Now, identify which is the right hyper plane to classify star and circle.

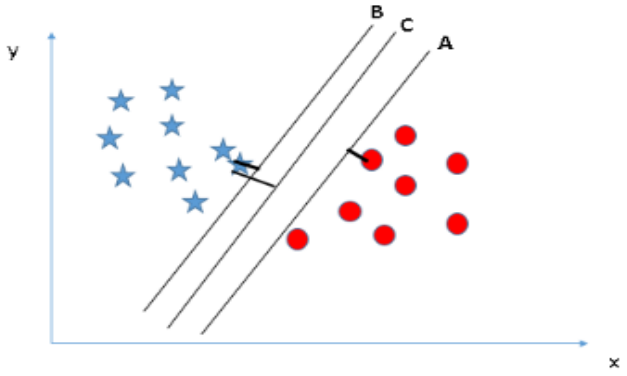


Here we need to remember a thumb rule to identify the right hyper plane: "Select the hyper plane which separates the two classes better". In this scenario, hyper plane "B" has excellently performed this job.

Scenario-2: Here, we have three hyper planes (A, B and C) and all are separating the classes well. Now, how can we identify the right hyper plane?

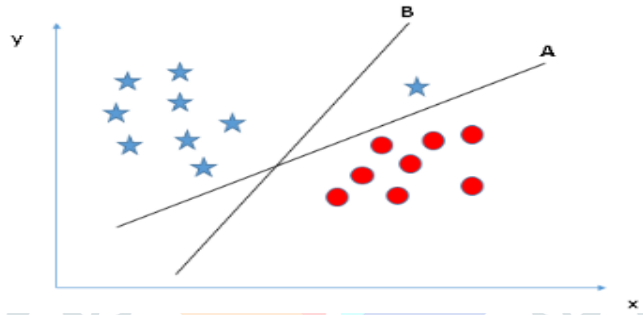


Here, maximizing the distances between nearest data points (either class) and hyper plane will help us to decide the right hyper plane. This distance is called as **Margin**.

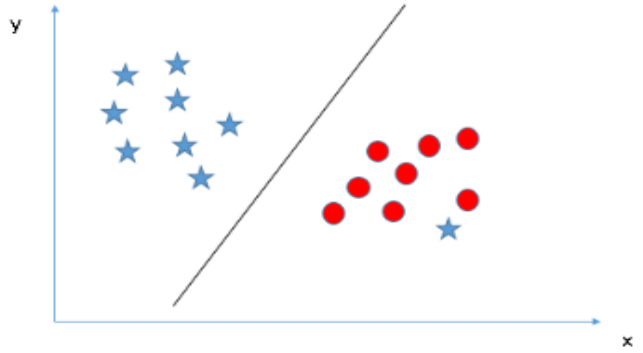


Above, you can see that the margin for hyper plane C is high as compared to both A and B. Hence, we name the right hyper plane as C. Another lightning reason for selecting the hyper plane with higher margin is robustness. If we select a hyper plane having low margin then there is high chance of miss-classification.

Scenario-3: Some of you may have selected the hyper plane B as it has higher margin compared to A. But, here is the catch; SVM selects the hyper plane which classifies the classes accurately prior to maximizing margin. Here, hyper plane B has a classification error and A has classified all correctly. Therefore, the right hyper plane is A.

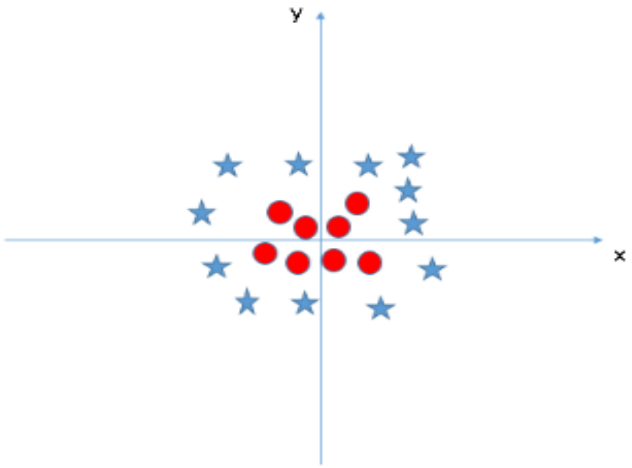


Scenario-4: Below, I am unable to separate the two classes using a straight line, as one of star lies in the territory of other (circle) class as an outlier.



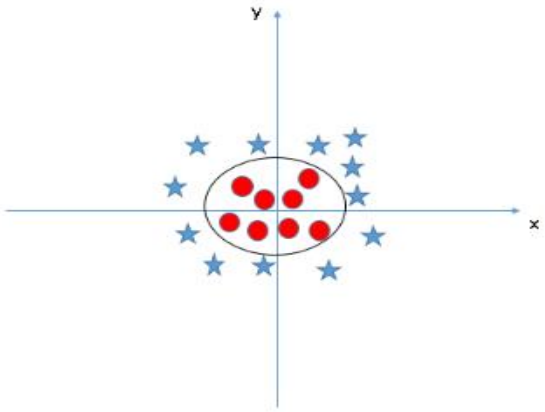
As I have already mentioned, one star at other end is like an outlier for star class. SVM has a powerful feature to avoid outliers and find the hyper-plane that has maximum margin. Hence, SVM is robust to outliers.

Scenario-5: In the scenario below, we can't have linear hyper plane between the two classes, so how does SVM classify these two classes? Upto this we have only looked at the linear hyper plane.



In SVM, it is very easy to have a linear hyper plane between two classes. But, another burning question is, should we need to add features manually to have a hyper plane. No, SVM has a special function called the **kernel**. This is a function which takes low dimensional input space and transforms it to a higher dimensional space. It converts non-separable problem to separable problem, these functions are called kernels. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then find out the process to separate the data based on the labels or outputs you have defined.

If we look at the hyper plane in original input space it looks like a circle:



IV. SVM APPLIED ON IDS

In the step of intrusion detection, the generated rules are used to classify the incoming network packets as follows:

KDD dataset:

The computer network intrusion detection systems use the KDD cup 1999 classification analysis of network traffic. The KDD stands for Knowledge Discovery and Data mining which is the most popular professional organization of data miners. The KDD organized the annual data mining knowledge discovery and data mining competition called KDD cup.

The KDD cup dataset consist of 41 features. Among the 41 features, 1-9 are used to represent basic feature of packet, 10-22 are the content features, 23-31 are used to represent traffic features and the 32-41 are host base features. Also the KDD cup comprises normal and 22 different types of attack. These attacks are mainly categorized into four classes are as follows:

- Denial of service
- Probing
- User to root
- Remote user

Feature Selection:

When presented data with very high random variables, models choke because **training time** increases exponentially with number of feature and models have increasing risk of **over fitting** with increases number of features.

The four major reasons to use feature selection are:

- It allows the machine learning algorithm to train faster.

- It reduces the complexity of a model and makes it easier to clarify.
- It improves the accuracy of a model.
- It reduces over fitting.

Accuracy depends on the number of fundamental feature in dataset and time complexity depends on the size of data. So we take all 41 features in intrusion detection system to improve the accuracy.

Feature Scaling:

Feature scaling is a method used to standardize the range of independent variables of data. The main advantage of feature scaling is to ignore attributes in greater numeric ranges dominating those in smaller numeric ranges.

Standardization involves rescaling the features. Standardization has the properties of a standard normal distribution with a mean of zero and a standard deviation of one.

$$Z = \frac{x - \mu}{\sigma}$$

Where μ is the mean of the training samples or zero if with_mean=False, and σ is the standard deviation of the training samples or one if with_std=False. Standardization compares features that have different units or scales. Here, standardizing tends to make the training process well behaved because the numerical condition of the optimization problems is improved.

V. DETECTION ALGORITHM ARCHITECTURE

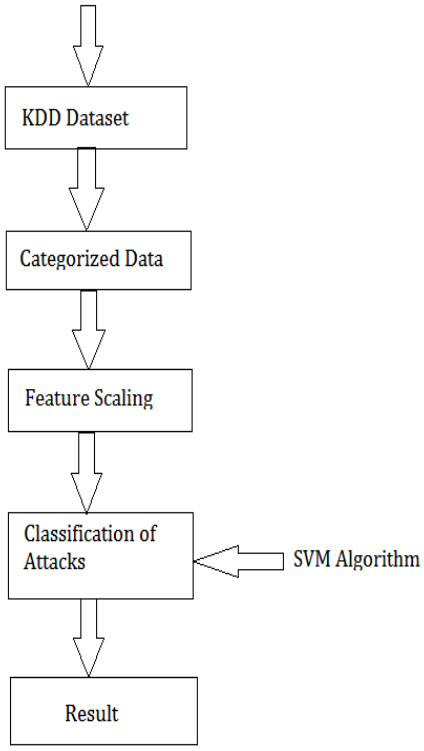


Fig. System Workflow

VI. KERNEL SVM

SVM algorithms use a set of mathematical functions which is the kernel. The kernel function is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel: linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

The kernel functions return the inner product between two points in a suitable feature space. Hence by defining a notion of similarity with the little computational cost even in very high dimensional space.

VII. IMPLEMENTATION AND RESULT

In machine learning algorithm, result can be represented in the form of confusion matrix, this matrix is also known as error matrix. It is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning. A confusion matrix is a summary of prediction results on a problem of classification. The number of correct and incorrect prediction are summarized with count values and broken down by each class. This is the key to the confusion matrix. Each row of the matrix represents the instances in a predicted class on the other hand column of the matrix represents the instances in an actual class or vice versa. The system gives the output in the form of confusion matrix is as follows:

Table 7.1: confusion matrix

	0	1	2	3	4
0	45847	75	5	0	0
1	48	66964	198	132	1
2	4	216	11436	0	0
3	4	151	0	840	0
4	1	25	1	2	23

In above matrix, each row of the instances represents machine predicted output and each column of instances represents the actual output. There are various types of attacks mainly categorized into normal, DOS, probe, R2L and U2R etc. Related to this matrix there four basic terms as follows:

- True positive:

If observation and prediction to be positive then it is true positive.

- True negative:

If observation and prediction to be negative then it is true negative.

- False positive:

If observation is negative and prediction is positive
Then it is false positive.

- False negative:

If observation is positive and prediction is negative then it is false negative

VIII. CONCLUSION

In this way we have implemented analysis of different features in dataset for this we have performed feature selection method based on dataset. This method reduces the time complexity and improves the accuracy. Then feature scaling is performed on continuous variables to limits the range of variables so that they can be compared on common ground. SVM classifiers perform classification of attacks and measure the accuracy between output of the dataset and predicted output of the machine. System gives the output in the form of confusion matrix; it shows the ways in which classification model is confused when it makes predictions.

.

REFERENCES

[1] Pavan Kaur and Dr.Dinesh Kumar "A study on intrusion detection based on KDD'99 benchmark dataset" IJERM (ISSN:2349- 2058, Volume-02, Issue-05 May2015)
[2] Adetunmbi A.Olusola., Adeola S.Oladele. and Daramola O.Abosede "Analysis of KDD'99 intrusion detection dataset for selection of relevance features" WCECS 2010, October 20-22, 2010, San Francisco, USA
[3] Nilotpal Chakraborty "Intrusion detection system and Intrusion prevention system: A comparative study" IJCBR (ISSN(online):2229-6166 ,Volume 4, Issue 2, May 2013)
[4] Dr. S.Vijayarani1 and Ms. Maria Sylvia's "Intrusion detection system-A study" IJSPTM, Vol 4,No1,February 2015
[5] https://www.analyticsvidhya.com
[6] https://www.kaggle.com