IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Anomaly Based Intrusion Detection System by Using Support Vector Machine in Network Traffic

1st A Srinivas , 2nd Dr.K Sagar Ph.D. 1st Asst professor,2ndProfessor Computer Science and Engineering Sree chaitanya College Of Engineering , KNR, india, Chaitanya Bharathi Institute of Technology, hyd. India

Abstract— In the network communication systems , network intrusion is the most important concern, because network attackers increased nowadays. to prevent such attacks by using intrusion detection tools and systems. Network attack is a devastating problem for network services. SVM has become one of the popular ML algorithm used for intrusion detection due to their good generalization nature and the ability to overcome the dimensionality problem, number of dimensions still affects the performance of SVM-based IDS. Machine learning is an effective analysis tool to detect any suspicious events occurred in the network traffic flow. In this paper, we developed a classifier model based on SVM based algorithms for network intrusion detection. The NSL-KDD dataset, a much improved version of the original KDDCUP'99 dataset, was used to evaluate the performance of our algorithm. The main task of our detection algorithm was to classify whether the incoming network traffics are normal or an attack, based on 41 features describing every pattern of network traffic. The detection accuracy 95 % was achieved using SVM.. The results of SVM visualized.

Keywords*— Network Intrusion, Support Vector Machine, accuracy, precision.*

I.INTRODUCTION

Network Security maintenance is one of the major safetyconcerns for neutralizing any unwanted activities. It is not only for protecting data and network privacy issues but also for avoiding any hazardous situations. For decades, Network security is one of the major issues and different types of developed systems are being implemented. Network intrusion is an unauthorized activity over the network that steals any important and classified data. Also sometimes it's the reason of unavailability of network services. The unexpected anomaly occurs frequently and a great loss to internet cyber world in terms of data security, the safety of potential information's etc. Therefore, the security system has to be robust, dependable and well configured. Traditionally, network intrusion detection systems (NIDS) are broadly classified based on the style of detection they are using: systems relying on *misuse-detection* monitor activity withprecise descriptions of known malicious behavior , while *anomaly-detection* systems have a notion of normal activity and flag deviations from that profile. Signature baseddetection system involves analyzing network traffic for a series of bytes or packet sequences known to be an anomaly Signature based type detection also has some disadvantages. A signature needs to be created for each attack and they are able to detect only those attacks. They are unable to detect any other novel attacks as their signatures are unknown to the detection scheme. Anomaly based NIDS operate based on theidea that the ambient traffic in a network collected over a period of time reflects the nature of the traffic that may be expected in the immediate future. Anomaly intrusion detection identifies deviations from the normal usage behavior patterns to identify the intrusion. The normal usage patterns are constructed from the statistical measures of the system features, for example, the CPU and I/O activities by a particular user or program. The behavior of the user is observed and any deviation from the constructed normal behavior is detected as intrusion.

2. LITERATURE REVIEW

Markov model in which the system being modeled is as- sumed to be a Markov process with unseen data. Prior research has shown that HMM analysis can be applied to identify particular kinds of malware (Annachhatre et al., 2015). In this technique, a Hidden Markov Model is trained against known malware features (e.g., operation code sequence) and once the training stage is completed, the trained model is applied to score the incoming traf- fic. The score is then contrasted to a predefined thresh- old, and a score greater than the threshold indicates malware. Likewise, if the score is less than the threshold, the traffic is identified as normal.

K-Nearest Neighbors (KNN) classifier: The k-Nearest Neighbor (k-NN) techniques is a typical non-parametric classifier applied in machine learning (Lin et al., 2015). The idea of these techniques is to name an unlabelled data sample to the class of its k nearest neighbors (where k is an integer defining the number of neigh- bours to be considered). Figure 5 illustrates a K-Nearest Neighbors classifier where k = 5. The point X represents an instance of unlabelled date which needs to be classi- fied. Amongst the five nearest neighbours of X there are three similar patterns from the class Intrusion and two from the class Normal. Taking a majority vote enables the assignment of X to the Intrusion class.

k-NN can be appropriately applied as a benchmark for all the other classifiers because it provides a good classi- fication performance in most IDSs (Lin et al., 2015).

AIDS based on machine learning techniques

Machine learning is the process of extracting knowledge from large quantities of data. Machine learning models com- prise of a set of rules, methods, or complex "transfer func- tions" that can be applied to find interesting data patterns, or to recognise or predict behaviour (Dua & Du, 2016).

Machine learning techniques have been applied exten- sively in the area of AIDS. Several algorithms and tech- niques such as clustering, neural networks, association rules, decision trees, genetic algorithms, and nearest neighbour methods, have been applied for discovering the knowledge from intrusion datasets (Kshetri & Voas, 2017; Xiao et al, 2018).

Some prior research has examined the use of different techniques to build AIDSs. Chebrolu et al. examined the performance of two feature selection algorithms involv- ing Bayesian networks (BN) and Classification Regres- sion Trees (CRC) and combined these methods for higher accuracy (Chebrolu et al., 2005).

Bajaj et al. proposed a technique for feature selection using a combination of feature selection algorithms such as Information Gain (IG) and Correlation Attribute evaluation. They tested the performance of the selected features by applying different classification algorithms such as C4.5, naïve Bayes, NB-Tree and Multi-Layer Per- ceptron (Khraisat et al., 2018; Bajaj & Arora, 2013). A genetic-fuzzy rule mining method has been used to

evaluate the importance of IDS features (Elhag et al., 2015). Thaseen et al. proposed NIDS by using Random Tree model to improve the accuracy and reduce the false alarm rate (Thaseen & Kumar, 2013). Subramanian et al. proposed classifying NSL-KDD dataset using decision tree algorithms to construct a model with respect to their metric data and studying the performance of deci- sion tree algorithms (Subramanian et al., 2012).

Comparison of various ML Algorithms used for IDS

In this paper survey of intrusion detection using ML algorithm has been presented and discussed.

Paper	Dataset	Detection	Infrastructure	Algorithm used	Evaluation	Outcomes
Goutham (et.al,2018	KDDCUP9 9	Intruder Detection	R programming and weka tool	Naïve Bayes Adaptive PART ensemble method	Prediction, Recall ,Accuracy	The result of the paper shows that the ensemble approach by bootstrapping achieves better performance than the other classifier.
Elsaeidy (et al, 2019)	Smart water distribution	DDoS attack	Java SDK 1.8, weka libraries, matlab 9.1	K-means, deep RBM, FFNN, automated FFNN, RF, SVM.	F-measures	The result of the paper shows that automated FFNN outperforms all other algorithm
Mehmood (et al, 2016)	KDDCUP9 9	Intruder detection like DoS, R2L, U2R.	.------	J48,Naïve Bayes, decision tables	True positive r ate, false positive rate, precision	The result of the paper shows that j.48 algorithm achieve better performance even under the redundant features among all other algorithm
Aburomma n (et al, 2016)	KDD-99	Intrusion detection	.----	PCA-LDA Ensemble classification	Overallaccuracy, Falsepositive, Falsenegative	The result of the paper show that ensemble approach LDA-

						PCA feature extraction is better than a single feature extraction algorithm, by having less false positive rate (0.0196).
Jan (et al, 2018)	Simulated dataset	DoS/DDoS attack	Matlab version 2018b simulation tool	SVM	Accuracy, True positive rate, False positive rate, False detection rate. DF, ANN, Logistic	The result of the paper achieves the light weight IDS for IoT. Experiments show that packet arrival rate and SVM classifier is enough to detect intrusions on IoT.
Hasan (et al, 2019	Kaggle	Attack and Anomaly detection	Framework used pandas, numpy, matplotlib, seaborn, scikitlearn, keras.	DF, ANN, Logistic Regression	accuracy, precision, recall, f1 score, ROC.	DF Is good technique to use in IoT for IDS with the accuracy of 99.4%.

Now a days, Machine learning techniques are heavily being adapted and developed in intrusion detection to enhance the efficacy of the systems [7] and in other applications as well [27. Suthaharan [8] in his work stated that due to the large size and redundant data in the datasets the computation cost of the machine learning methods increases drastically. They proposed ellipsoid-based technique which detects anomalies and side by side cleans the dataset. The research of [9] deals with intrusion detection technique which is a combination of k means clustering, neuro-fuzzy logic techniques. and radial basis support vector machine. In their technique, firstly k-means clustering is used to spawn the training subsets, on them various neuro fuzzy models are trained, after that KNN classification is generated and finally classification task is carried by KNN technique.

We propose a method that is based on the classification algorithm named as SVM and use it todetect the intrusions. number of samples is more [10]. we present a model that we implemented an intrusion detection system for classification of intrusion types which outperforms the support vector machine method and the nearest centroid classification method in terms of accuracy, the detection rate and falsealarm. An analysis has been performed for each type of attack mentioned in the dataset that has been utilized for this study

3.NSL-KDD Dataset

The dataset to be used in this research is the NSL-KDD dataset [11] which is a new dataset for the evaluation of researches in network intrusion detection system. It consistsof selected records of the complete KDD 99 dataset. NSL- KDD dataset solve the issues of KDD 99 benchmark and connection record contains 41 features. Among the 41 features, 34 features are numeric and 7 features are symbolic or discrete. The NSL-KDD training set contains a total of 22 training attack types; with an additional 17 typesin the testing set only. Table I gives the description of NSL-KDD Dataset Features.

Table I: Description of NSL-KDD Dataset Features

Feature name	Variable type	Description
Duration	C	No. of seconds of the connection
Protocol_type	D	Type of protocol Eg.TCP,UDP ,ICMP
Service	D	Network service on the destination eg:http,telnet,etc
Flag	D	Normal or error status of the connection
src_bytes	C	Number of data bytes from source to destination
dst_bytes	C	Number of data bytes from destination to source
Land	D	1-connection is from the same host/port: 0-otherwise
Wrong_fragment	C	No. of 'wrong' fragments
Urgent	C	No of urgent fragments
Hot	C	The count of access to system directories, creation and execution of programs
Num_failed_logins	C	No. of failed login attempts
Logged_in	D	1-successfully logged in 0-otherwise
num_compromised	C	No. of compromised conditions
Root_shell	C	1-root shell is obtained;0 otherwise
Su_attempted	C	1-'su root' command attempted;0 otherwise
Num_root	C	No .of root accesses
num_file_creations	C	Number of file creation operations
Num_shells	C	No of shell prompts
Num_access_files	C	No. of write ,delete and create operations on access control files
Num_outbound_cmds	C	No. of outbound commands in an ftp session
Is_hot_login	D	1-the login belongs to the 'hot' list 0: otherwise
Count	C	No. of connections to the same host as the current connection in the past seconds
Srv_count	C	No of connections to the same host as the current connection in the past 2 seconds
serror_rate	C	% of connections that have 'SYN' errors to the same host

Srv_serror_rate	C	% of connections that have 'SYN' errors to the same service
Rerror_rate	C	% of connections that have 'REJ' errors to the same host
Srv_diff_host_rate	C	% of connections to different services and to the same host
Dst_host_count	C	No of connections to the same host to the destination host as the current connection in the past 2 seconds
Dst_host_srv_count	C	No of connections from the same service to the destination host as the current connection in the past 2 seconds
dst_host_srv_count	C	No. of connections from the same service to the destination host as the current connection in the past 2 seconds
Dst_host_srv_count	C	No. of connections from the same service to the destination host as the current connection in the past 2 seconds
Dst_host_same_srv_rate	C	% of connections from the same service to the destination host
Dst_host_diff_srv_rate	C	% of connections from the different services to the destination host
Dst_host_same_src_port_rate	C	% of connections from the port services to the destination host
Dst_host_srv_diff_host_rate	C	% of connections from the different hosts from the same service to destination host
Dst_host_serror_rate	C	% of connections that have 'SYN" errors to same host to the destination host
dst_host_srv_serror_rate	C	% of connections that have 'SYN' errors from the same service to the destination host
Dst_host_rerror_rate	C	% of connections that have 'REJ' errors from the same host to destination host
Dst_host_srv_rerror_rate	C	% of connections that have 'REJ' errors from the same service to the destination host

NSL – KDD Dataset Preprocessing:
Classification algorithms are not able to process NSL - KDDdataset in its current format.
Hence we need to preprocess the datasets before training themodel.
Preprocessing contains below steps:

- Mapping symbolic features to numeric value.
- Implementing scaling since the data have significantly varying resolution and ranges. The attribute data are scaled to fall within the range [-1, 1].
- Attack names were mapped to one of the two classes, 0 for Normal, 1 for Attack.

Missing values in data

Types of Network Attacks:

Identify the Type	Meaning	Specific Classification Identification
Normal	Normal record	Normal
DOS	Denial of service attacks	Neptune,pod,land, back,smurf, teardrop
Probing	Monitoring and other exploration activities	Ipsweep,nmap,portsweep ,satan etc.
R2L	Unauthorized access from remote machine	Imap,ftp_write, Warezclient,multihop, phf,spy,guess_passwd, warezmast
U2R	Unauthorized access to local super user privileges by ordinary users	Loadmodule, buffer_overflow,rootkit, per

4. Classification Model:

In general, the category of problems which contains data as well as the additional attributes that we want to predict comesunder supervised learning approach. Under supervised learning approach the classification problem comes into account into when the instances belong to two or more classes and our intention is forecast

the unlabeled instances under the procedure of supervised learning methods. By using SVM classification method this method best suited for high dimensional spaces. it utilizes subset of training data points in the decision function called as support vectors, also it is adroit as for the decision function various kinds of kernel functions can be stated . If the count of features is bigger than the count of samples this technique is liable to give mediocreperformance.

A) Support Vector Machine:

The SVM uses a portion of the data to train the system, finding several support vectors that represent the training data. These support vectors will form a SVM model. A basic input data format and output data domains are listed as follows

(Xi, Yi)...................Xn, Yn)

Where

$X \in R^m$ and $Y \in \{0, 1\}$

(Xi, Yi) .... (Xn, Yn) is training data records, n is the numbers of samples m is the inputs vector, and y belongs to category of class ‘0’ or class ’1’ respectively. On theproblem of linear, a hyper plane can be divided into the two categories as shown in Figure.

The hyper plan formula is:

$(w \cdot x) + b = 0$The category formula is:

$(w. x) + b \geq 0$ if $Yi = 1$$(w. x) + b \leq 0$ if $Yi = 0$

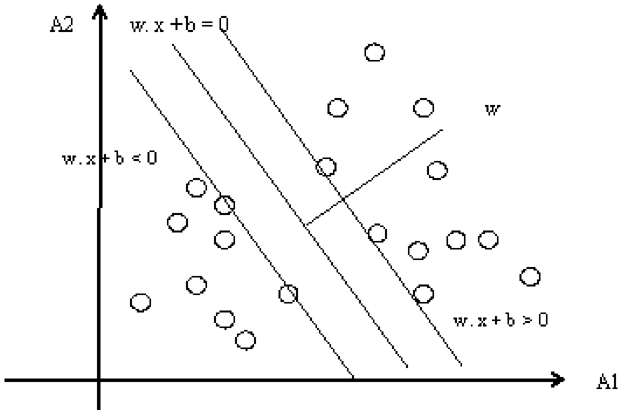

Figure 1-Classifiaction using of SVM

A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one “target value" (class labels: Normal or Attack) and several “attributes" (features).The goal of SVM is to produce a model which predicts target value of data instance in the testing set which is given only attributes. To attain this goal there are four different kernel functions.in this experiment RBF kernel function is used

The main advantage of the kernel methods is the possibility of using linear models in a nonlinear subspace by an implicit transformation of patterns to a high-dimensional feature space without computing their images directly. An appropriately constructed kernel results in a model that fits well to the structure underlying data and doesn’t over-fit to the sample. is verified. Alternative evaluation measures that outperform presented methods are proposed. Optimization leveraging these measures results in parameters corresponding to the classifiers that achieve minimal error rate for RBF kernel.
The Formula for RBF Kernel Optimization function :

$$K(X, X') = \exp(-\|X_i - X'_j\|2/2\ \sigma 2)$$

$$\begin{aligned}
\exp\left(-\frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|^2\right) &= \exp\left(\frac{2}{2}\mathbf{x}^\top\mathbf{x}' - \frac{1}{2}\|\mathbf{x}\|^2 - \frac{1}{2}\|\mathbf{x}'\|^2\right) \\
&= \exp(\mathbf{x}^\top\mathbf{x}')\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)\exp\left(-\frac{1}{2}\|\mathbf{x}'\|^2\right) \\
&= \sum_{j=0}^{\infty}\frac{(\mathbf{x}^\top\mathbf{x}')^j}{j!}\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)\exp\left(-\frac{1}{2}\|\mathbf{x}'\|^2\right) \\
&= \sum_{j=0}^{\infty}\sum_{\sum n_i = j}\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)\frac{x_1^{n_1}\cdots x_k^{n_k}}{\sqrt{n_1!\cdots n_k!}}\exp\left(-\frac{1}{2}\|\mathbf{x}'\|^2\right)\frac{x_1'^{n_1}\cdots x_k'^{n_k}}{\sqrt{n_1!\cdots n_k!}}
\end{aligned}$$

5. Result and Discussion

The performance of all the classifiers was computed by utilizing a matrix known as confusion matrix. It is a standard metric for benchmarking the effectiveness and robustness of a classification algorithm. Using the confusion matrix, measures like accuracy, detection rate and false alarm rate have been computed which are the generic criteria for evaluating the performance of the IDS. These metrics have been utilized in a number of studies and they ensure a viable means of deciding the efficiency of the model for detecting the intrusions within systems. For a decent level of performance, the intrusion detection system (IDS) needs high accuracy and precision and conversely false alarm rate should be low. These terms are given by the following formulae:

Accuracy = (TP+TN) / (TP+TN+FP+FN)

Precision = (TP) / (TP+TN)

True positive rate (TPR) = (TP) / (TP+TN)

False positive rate (FPR) = (FP) / (TN+FP)

True negative rate (TNR) = (TN) / (TP + FN)

False negative rate = (FN) / TP+FN)

Following figure represents a matrix known as confusion matrix. True positive (TP) indicates the number of instances having the class label of attack and were correctly classified as an attack. True negative (TN) indicates the number ofinstances having the class label of normal and were correctly classified as normal. False positive (FP) indicates the number of instances that have a label of being valid but have been incorrectly classified as intrusion. False negative (FN)indicates the number of instances that were having a label of intrusion but were incorrectly classified as normal by theIDS.

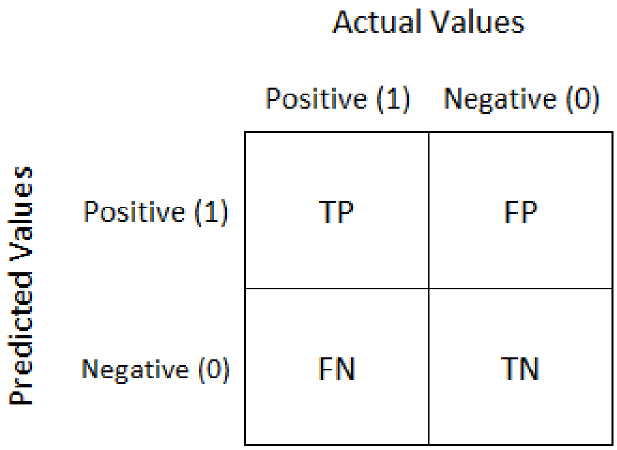

Fig. Confusion matrix

Experimental Analysis:

Following figure shows the prediction result of SVM: method

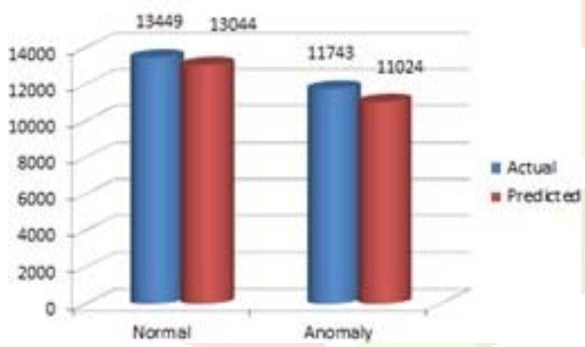

Figure 2- Prediction Result of Support Vector Machine

Table II: Result of SVM tclassification model based on performance measure.

Algorithm	True Positive rate	False Positive rate	True Negative rate	False Negative rate
SVM	93.87%	3.01%	96.98%	6.12%

Algorithm	Accuracy	Precision
SVM	95.53 %	96.45%

6. CONCLUSION

In this paper, we have scrutinized some new techniques for intrusion detection and evaluated their performance based on the benchmark KDD Cup 99 Intrusion data. An Intrusion Detection System that was able to assay the dynamic and complex nature of intrusion activities has been built.

The performances of the different kernel based approaches have been observed on the basis of their accuracy, false negative rate and precision. The results indicate that the ability of the SVM classification depends mainly on the kernel type and the setting of the parameters. Research in intrusion detection using SVM approach is still an ongoing area due to good performance. By using SVM way in order to maximize the performance rate and minimize the false negative rate.

7. REFERENCES

[1] Annachhatre, T. H. Austin, and M. Stamp, "Hidden Markov models for malware classification," Journal of Computer Virology and Hacking Techniques, vol. 11, no. 2, pp. 59–73, 2015/05/01 2015

[2] Lin, S.-W. Ke, and C.-F. Tsai, "CANN: an intrusion detection system based on combining cluster centers and nearest neighbors," Knowl-Based Syst, vol. 78, no. Supplement C, pp. 13–21, 2015/04/01/ 2015

[3] Kshetri N, VoasJ (2017) Hacking power grids: a current problem. Computer50(12):91–95

[4] Khraisat A, Gondal I, Vamplew P (2018) An anomaly intrusion detection system using C5 decision tree classifier. In: Trends and applications in knowledge discovery and data mining. Springer International Publishing, Cham, pp 149–155

[5] S. Elhag, A. Fernández, A. Bawakid, S. Alshomrani, and F. Herrera, "On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on intrusion detection systems," Expert Syst Appl, vol. 42, no. 1, pp. 193–202, 1// 2015

[6] Khraisat et al. Cybersecurity (2019) 2:20 https://doi.org/10.1186/s42400-019-0038-7

[7] S. S. Roy, V. M. Viswanatham - Classifying Spam Emails Using Artificial Intelligent Techniques. In International Journal of Engineering Research in Africa, vol. 22, pp. 152-161. Trans Tech Publications, 2016.

[8] S. S. Roy, D. Mittal, A. Basu, A. Abraham - Stock Market Forecasting Using LASSO Linear Regression Model. In AfroEuropean Conference for Industrial Advancement, pp. 371-381. Springer International Publishing, 2015.

[9] S. Suthaharan - An iterative ellipsoid-based anomaly detection technique for intrusion detection systems, In Southeast on, Proceedings of IEEE, pp. 1-6, 2012.

[10] A. Chandrasekhar, K. Raghuveer - Intrusion detection technique by using k-means, fuzzy neural network and svm classifiers, In Computer Communication and Informatics (ICCCI), 2013 International Conference, pp. 1-7. [5] S. Adusumilli, D. Bhatt, H. Wang, V. Devabhaktuni, P. Bhattacharya - A novel hybrid approach utilizing principal component regression and random forest regression to bridge the eripod of GPS outages, Neurocomputing, 2015.

[11] J. Ali, R. Khan, N. Ahmad, I. Maqsood - Random forests and decision trees, IJCSI International Journal of Computer Science Issues, vol. 9, no. 5, 2012.