INTERNATIONAL JOURNAL OF TECHO-ENGINEERING
ISSN : 2057-5688

MACHINE LEARNING TECHNIQUE USING NETWORK INTRUSION DETECTION WITH FEATURE SELECTION

Kunduru Sai Asritha[1], Neelam Lahre[2], Thimmaiahgari Pranay[3], Dasari Gokul Pavan[4] , Dr.A.Anjaiah[5]

[1,2,3,4] UG Scholars, Department of CSE, ***ST.PETER'S ENGINEERING COLLEGE***, Hyderabad, Telangana, India.
[5]Associate Professor, Department of CSE, ***ST.PETER'S ENGINEERING COLLEGE***, Hyderabad, Telangana, India.

ABSTRACT:

In this paper author is evaluating performance of two supervised machine learning algorithms such as SVM (Support Vector Machine) and ANN (Artificial Neural Networks). Machine learning algorithms will be used to detect whether request data contains normal or attack (anomaly) signatures. Now-a-days all services are available on internet and malicious users can attack client or server machines through this internet and to avoid such attack request IDS (Network Intrusion Detection System) will be used, IDS will monitor request data and then check if its contains normal or attack signatures, if contains attack signatures then request will be dropped. IDS will be trained with all possible attacks signatures with machine learning algorithms and then generate train model, whenever new request signatures arrived then this model applied on new request to determine whether it contains normal or attack signatures. In this paper we are evaluating performance of two machine learning algorithms such as SVM and ANN and through experiment we conclude that ANN outperform existing SVM in terms of accuracy. To avoid all attacks IDS systems has developed which process each incoming request to detect such attacks and if request is coming from genuine users then only it will forward to server for processing, if request contains attack signatures then IDS will drop that request and log such request data into dataset for future detection purpose. To detect such attacks IDS will be prior train with all possible attacks signatures coming from malicious user's request and then generate a training model. Upon receiving new request IDS will apply that request on that train model to predict it class whether request belongs to normal class or attack class. To train such models and prediction various data mining classification or prediction algorithms will be used. In this paper author is evaluating performance of SVM and ANN. In this algorithms author has applied Correlation Based and Chi-Square Based feature selection algorithms to reduce dataset size, this feature selection algorithms removed irrelevant data from dataset and then used model with important features, due to this features selection algorithms dataset size will reduce and accuracy of prediction will increase.

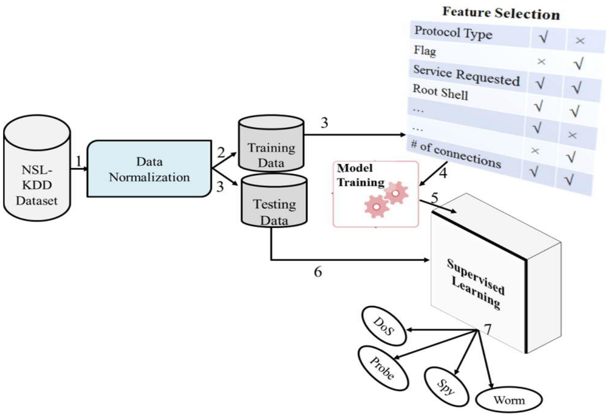
Keyword:

INTRODUCTION

With the wide spreading usages of internet and increases in access to online contents, cybercrime is also happening at an increasing rate [1-2]. Intrusion detection is the first step to prevent security attack. Hence the security solutions such as Firewall, Intrusion Detection System (IDS), Unified Threat Modeling (UTM) and Intrusion Prevention System (IPS) are getting much attention in studies. IDS detects attacks from a variety of systems and network sources by collecting information and then analyzes the information for possible security breaches [3]. The network based IDS analyzes the data packets that travel over a network and this analysis are carried out in two ways. Till today anomaly based detection is far behind than the detection that works based on signature and hence anomaly based detection still remains a major area for research [4-5]. The challenges with anomaly based intrusion detection are that it needs to deal with novel attack for which there is no prior knowledge to identify the anomaly. Hence the system somehow needs to have the intelligence to segregate which traffic is harmless and which one is malicious or anomalous and for that machine learning techniques are being explored by the researchers over the last few years [6]. IDS however is not an answer to all security related problems. For example, IDS cannot compensate weak identification and authentication mechanisms or if there is a weakness in the network protocols

Studying the field of intrusion detection first started in 1980 and the first such model was published in 1987 [7]. For the last few decades, though huge commercial investments and substantial research were done, intrusion detection technology is still immature and hence not effective [7]. While network IDS that works based on signature have seen commercial success and widespread adoption by the technology based organization throughout the globe, anomaly based network IDS have not gained success in the same scale. Due to that reason in the field of IDS, currently anomaly based detection is a major focus area of research and development [8]. And before going to any wide scale deployment of anomaly based intrusion detection system, key issues remain to be solved [8]. But the literature today is limited when it comes to compare on how intrusion detection performs when using supervised machine learning techniques [9]. To protect target systems and networks against malicious activities anomaly-based network IDS is a valuable technology. Despite the variety of anomaly-based network intrusion detection techniques described in the literature in recent years [8], anomaly detection functionalities enabled security tools are just beginning to appear, and some important problems remain to be solved. Several anomaly based techniques have been proposed including Linear Regression, Support Vector Machines (SVM), Genetic Algorithm, Gaussian mixture model, knearest neighbor algorithm, Naive Bayes classifier, Decision Tree [3,5]. Among them the most widely used learning algorithm is SVM as it has already established itself on different types of problem [10]. One major issue on anomaly based detection is though all these proposed techniques can detect novel attacks but they all suffer a high false alarm rate in general. The cause behind is the complexity of generating profiles of practical normal behavior by learning from the training data sets [11]. Today Artificial Neural Network (ANN) are often trained by the back propagation algorithm, which had been around since 1970 as the reverse mode of automatic differentiation [12].

The major challenges in evaluating performance of network IDS is the unavailability of a comprehensive network based data set [13]. Most of the proposed anomaly based techniques found in the literature were evaluated using KDD CUP 99 dataset [14]. In this paper we used SVM and ANN –two machine learning techniques, on NSLKDD [15] which is a popular benchmark dataset for network intrusion



LITERATURE SURVEY

1. “A macro-social exploratory analysis of the rate of interstate cyber-victimization:

This study examines whether macro-level opportunity indicators affect cyber-theft victimization. Based on the arguments from criminal opportunity theory, exposure to risk is measured by state-level patterns of internet access (where users access the internet). Other structural characteristics of states were measured to determine if variation in social structure impacted cyber-victimization across states. The current study found that structural conditions such as unemployment and non-urban population are associated with where users access the internet. Also, this study found that the proportion of users who access the internet only at home was positively associated with state-level counts of cyber-theft victimization. The theoretical implications of these findings are discussed.

2. Incremental anomaly-based intrusion detection system using limited labeled data:

With the proliferation of the internet and increased global access to online media, cybercrime is also occurring at an increasing rate. Currently, both personal users and companies are vulnerable to cybercrime. A number of tools including firewalls and Intrusion Detection Systems (IDS) can be used as defense mechanisms. A firewall acts as a checkpoint which allows packets to pass through according to predetermined conditions. In extreme cases, it may even disconnect all network traffic. An IDS, on the other hand, automates the monitoring process in computer networks. The streaming nature of data in computer networks poses a significant challenge in building IDS. In this paper, a method is proposed to overcome this problem by performing online classification on datasets. In doing so, an incremental naive Bayesian classifier is employed. Furthermore, active learning enables solving the problem using a small set of labeled data points which are often very expensive to acquire. The proposed method includes two groups of actions i.e. offline and online. The former involves data preprocessing while the latter introduces the NADAL online method. The proposed method is compared to the incremental naive Bayesian classifier using the NSL-KDD standard dataset. There are three advantages with the proposed method: (1) overcoming the streaming data challenge; (2) reducing the high cost associated with instance labeling; and (3) improved accuracy and Kappa compared to the incremental naive Bayesian approach. Thus, the method is well-suited to IDS applications.

3. Modeling and implementation approach to evaluate the intrusion detection system:

Intrusions detection systems (IDSs) are systems that try to detect attacks as they occur or when they were over. Research in this area had two objectives: first, reducing the impact of attacks; and secondly the evaluation of the system IDS. Indeed, in one hand the IDSs collect network traffic information from some sources present in the network or the computer system and then use these data to enhance the systems safety. In the other hand, the evaluation of IDS is a critical task. In fact, its important to note the difference between evaluating the effectiveness of an entire system and evaluating the characteristics of the system components. In this paper, we present an approach for IDS evaluating based on measuring the performance of its components. First of all, in order to implement the IDS SNORT components safely we have proposed a hardware platform based on embedded systems. Then we have tested it by using a generator of traffics and attacks based on Linux KALI (Backtrack) and Metasploite 3 Framework. The obtained results show that the IDS performance is closely related to the characteristics of these components

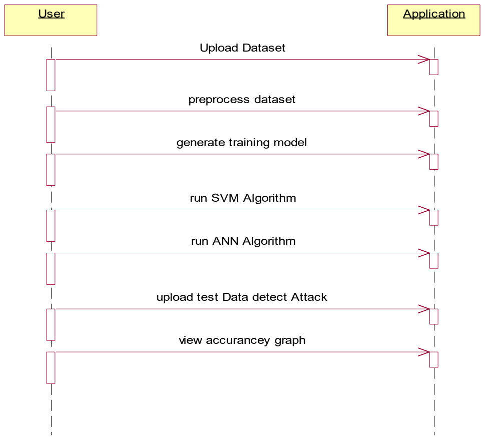
EXISTING SYSTEM

The major challenges in evaluating performance of network IDS is the unavailability of a comprehensive network based data set [13]. Most of the proposed anomaly based techniques found in the literature were evaluated using KDD CUP 99 dataset . In this paper we used SVM and ANN –two machine learning techniques, on NSLKDD which is a popular benchmark dataset for network intrusion.

The promise and the contribution machine learning did till today are fascinating. There are many real life applications we are using today offered by machine learning. It seems that machine learning will rule the world in coming days. Hence we came out into a hypothesis that the challenge of identifying new attacks or zero day attacks facing by the technology enabled organizations today can be overcome using machine learning techniques. Here we developed a supervised machine learning model that can classify unseen network traffic based on what is learnt from the seen traffic. We used both SVM and ANN learning algorithm to find the best classifier with higher accuracy and success rate

PROPOSED SYSTEM

The system proposed is composed of feature selection and learning algorithm show in Fig.1. Feature selection component are responsible to extract most relevant features or attributes to identify the instance to a particular group or class. The learning algorithm component builds the necessary intelligence or knowledge using the result found from the feature selection component. Using the training dataset, the model gets trained and builds its intelligence. Then the learned intelligences are applied to the testing dataset to measure the accuracy of home much the model correctly classified on unseen data.



Sequence Graph

MODULES

1. Feature Selection:

Feature selection is an important part in machine learning to reduce data dimensionality and extensive research carried out for a reliable feature selection method. For feature selection filter method and wrapper method have been used. In filter method, features are selected on the basis of their scores in various statistical tests that measure the relevance of features by their correlation with dependent variable or outcome variable. Wrapper method finds a subset of features by measuring the usefulness of a subset of feature with the dependent variable. Hence filter methods are independent of any machine learning algorithm whereas in wrapper method the best feature subset selected depends on the machine learning algorithm used to train the model. In wrapper method a subset evaluator uses all possible subsets and then uses a classification algorithm to convince classifiers from the features in each subset. The classifier consider the subset of feature with which the classification algorithm performs the best. To find the subset, the evaluator uses different search techniques like depth first search, random search, breadth first search or hybrid search. The filter method uses an attribute evaluator along with a ranker to rank all the features in the dataset. Here one feature is omitted at a time that has lower ranks and then sees the predictive accuracy of the classification algorithm. Weights or rank put by the ranker algorithms are different than those by the classification algorithm. Wrapper method is useful for machine learning test whereas filter method is suitable for data mining test because data mining has thousands of millions of features

2. Building Machine Intelligence:

Based on the best features found in the feature selection process, learning models are developed. To develop the learning model, machine learning algorithm is used. Training dataset is used to train the algorithm with the selected features. In supervised machine learning, each instance in the training dataset has the class it belongs to

3. Support Vector Machine (SVM) :

In SVM a separating hyper plane defines the classifier depending on the type of problem and available datasets. In case where dataset is one dimensional, the hyper plane is a point, for two dimensional data it is a separating line as shown in Fig 2, for three dimensional dataset, it is a plane and if the data dimension is higher it is a hyper plane. For a linearly separable dataset, the classifier or the decision function will have the form.

Artificial Neural Network (ANN)

Artificial Neural Network is another tool used in machine learning. As it name suggests, ANN is a system inspired by human brain system and replicate the learning system of human brain. It consists of input and output layers with one or more hidden layers in most cases as shown in Fig 3. The ANN uses a technique called back propagation to adjust the outcome with the expected result or class

SCREEN SHOTS

In this paper author is evaluating performance of two supervised machine learning algorithms such as SVM (Support Vector Machine) and ANN (Artificial Neural Networks). Machine learning algorithms will be used to detect whether request data contains normal or attack (anomaly) signatures. Now-a-days all services are available on internet and malicious users can attack client or server machines through this internet and to avoid such attack request IDS (Network Intrusion Detection System) will be used, IDS will monitor request data and then check if its contains normal or attack signatures, if contains attack signatures then request will be dropped

IDS will be trained with all possible attacks signatures with machine learning algorithms and then generate train model, whenever new request signatures arrived then this model applied on new request to determine whether it contains normal or attack signatures. In this paper we are evaluating performance of two machine learning algorithms such as SVM and ANN and through experiment we conclude that ANN outperform existing SVM in terms of accuracy

To avoid all attacks IDS systems has developed which process each incoming request to detect such attacks and if request is coming from genuine users then only it will forward to server for processing, if request contains attack signatures then IDS will drop that request and log such request data into dataset for future detection purpose.

To detect such attacks IDS will be prior train with all possible attacks signatures coming from malicious user’s request and then generate a training model. Upon receiving new request IDS will apply that request on that train model to predict it class whether request belongs to normal class or attack class. To train such models and prediction various data mining classification or prediction algorithms will be used.

In this paper author is evaluating performance of SVM and ANN.

In this algorithms author has applied Correlation Based and Chi-Square Based feature selection algorithms to reduce dataset size, this feature selection algorithms removed irrelevant data from dataset and then used model with important features, due to this features selection algorithms dataset size will reduce and accuracy of prediction will increase

To conduct experiment author has used NSL KDD Dataset and below is some example records of that dataset which contains request signatures. I have also used same dataset and this dataset is available inside ‘dataset’ folder.

Dataset example:

duration,protocol_type,service,flag,src_bytes,dst_bytes,land,wrong_fragment,urgent,hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files,num_outbound_cmds,is_host_login,is_guest_login,count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate, dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate,label

All above comma separated names in bold format are the names of request signature

0,tcp,ftp_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0,0,0,0,1,0,0,150,25,0.17,0.03,0.17,0,0,0,0.05,0,normal

0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,166,9,1,1,0,0,0.05,0.06,0,255,9,0.04,0.05,0,0,1,1,0,0,anamoly

Above two records are the signature values and last value contains class label such as normal request signature or attack signature. In second record ‘Neptune’ is a name of attack. Similarly in dataset you can find nearly 30 different names of attacks.

In above dataset records we can see some values are in string format such as tcp, ftp_data and these values are not important for prediction and these values will be remove out by applying PREPROCESSING Concept.

All attack names will not be identified by algorithm if it's given in string format so we need to assign numeric value for each attack. All this will be done in PREPROCESS steps and then new file will be generated called 'clean.txt' which will use to generate training model

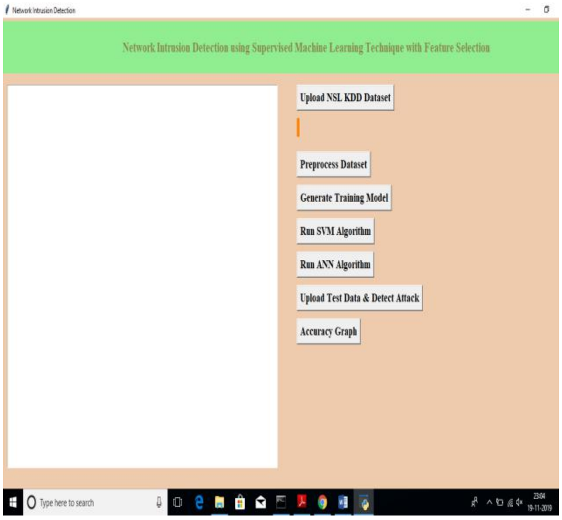
In below line i am assigning numeric id to each attack

"normal":0,"anamoly":1

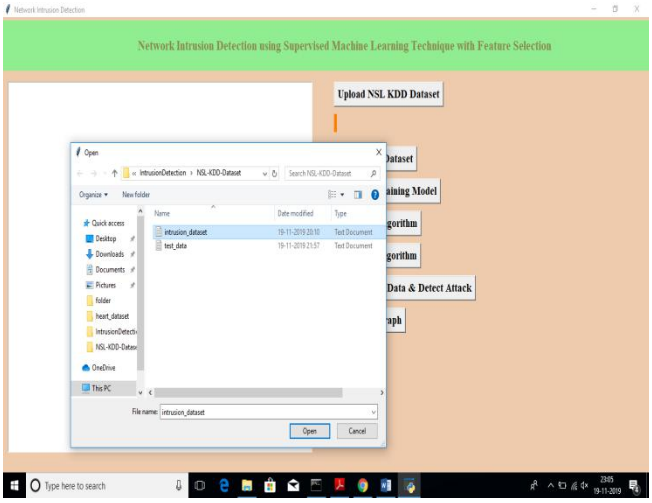
In above lines we can see normal is having id 0 and Anomaly has id 1 and goes on for all attacks

Before running code execute below two commands

Screen shots
Double click on 'run.bat' file to get below screen

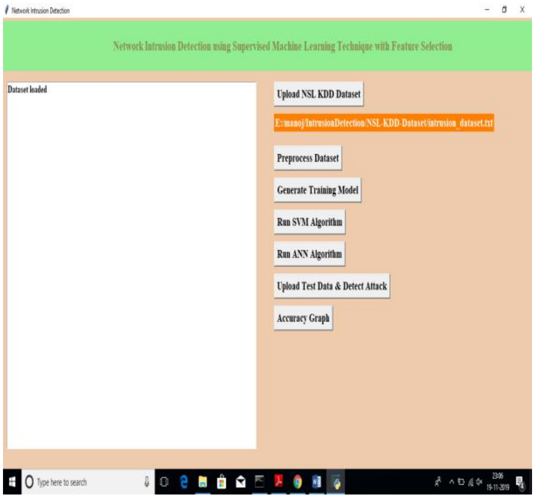

Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search

In above screen click on 'Upload NSL KDD Dataset' button and upload dataset



File name: intrusion_dataset

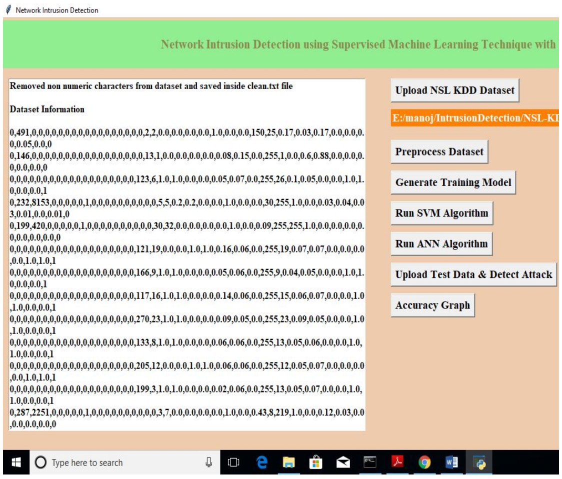
In above screen I am uploading 'intrusion_dataset.txt' file, after uploading dataset will get below screen


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search

Dataset loaded

E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_dataset.txt

Now click on 'Pre-process Dataset' button to clean dataset to remove string values from dataset and to convert attack names to numeric values

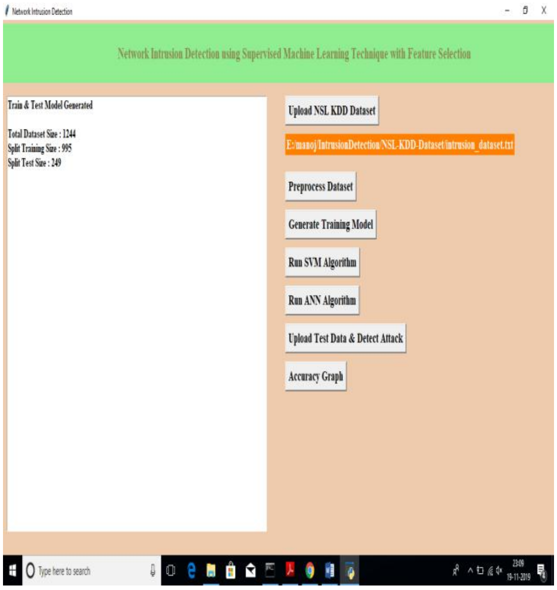

Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search

E:/manoj/IntrusionDetection/NSL-KI

```
Removed non numeric characters from dataset and saved inside clean.txt file

Dataset Information

0,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.0,0.0,0.0,0.0,1.0,0.0,0.0,150,25,0.17,0.03,0.17,0.0,0.0,0.
0,0.05,0.0,0
0,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,13,1,0.0,0.0,0.0,0.0,0.08,0.15,0.0,255,1,0.0,0.6,0.88,0.0,0.0,0.
0,0.0,0.0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,123,6,1.0,1.0,0.0,0.0,0.05,0.07,0.0,255,26,0.1,0.05,0.0,0.0,1.0,1.
0,0.0,0.0,1
0,232,8153,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,5,5,0.2,0.2,0.0,0.0,1.0,0.0,0.0,30,255,1.0,0.0,0.03,0.04,0.0
3,0.01,0.0,0.01,0
0,199,420,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,30,32,0.0,0.0,0.0,0.0,1.0,0.0,0.09,255,255,1.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,121,19,0.0,0.0,1.0,1.0,0.16,0.06,0.0,255,19,0.07,0.07,0.0,0.0,0.0
,0.0,1.0,1.0,1
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,166,9,1.0,1.0,0.0,0.0,0.05,0.06,0.0,255,9,0.04,0.05,0.0,0.0,1.0,1.
0,0.0,0.0,1
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,117,16,1.0,1.0,0.0,0.0,0.14,0.06,0.0,255,15,0.06,0.07,0.0,0.0,1.0
,1.0,0.0,0.0,1
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,270,23,1.0,1.0,0.0,0.0,0.09,0.05,0.0,255,23,0.09,0.05,0.0,0.0,1.0
,1.0,0.0,0.0,1
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,133,8,1.0,1.0,0.0,0.0,0.06,0.06,0.0,255,13,0.05,0.06,0.0,0.0,1.0,
1.0,0.0,0.0,1
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,205,12,0.0,0.0,1.0,1.0,0.06,0.06,0.0,255,12,0.05,0.07,0.0,0.0,0.0
,0.0,1.0,1.0,1
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,199,3,1.0,1.0,0.0,0.0,0.02,0.06,0.0,255,13,0.05,0.07,0.0,0.0,1.0,
1.0,0.0,0.0,1
0,287,2251,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,3,7,0.0,0.0,0.0,0.0,1.0,0.0,0.43,8,219,1.0,0.0,0.12,0.03,0.0
,0.0,0.0,0.0,0
```

After pre-processing all string values removed and convert string attack names to numeric values such as normal signature contains id 0 and anomaly attack contains signature id 1.

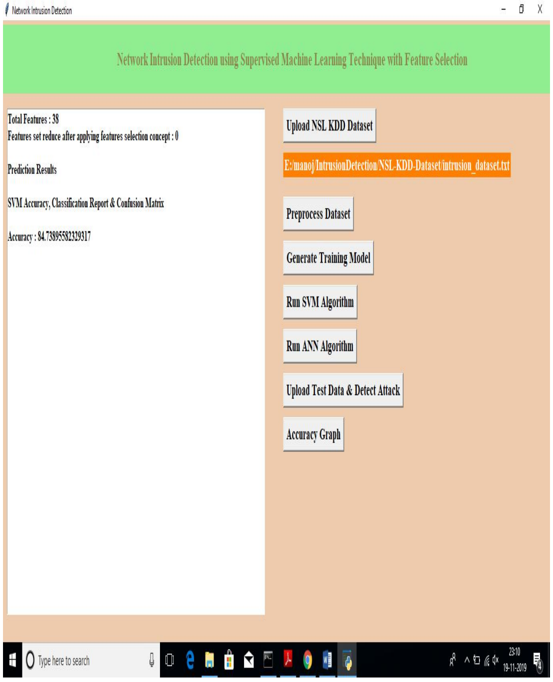
Now click on 'Generate Training Model' to split train and test data to generate model for prediction using SVM and ANN


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search
23:08
19-11-2019

E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_dataset.txt

```
Train & Test Model Generated

Total Dataset Size : 1244
Split Training Size : 995
Split Test Size : 249
```

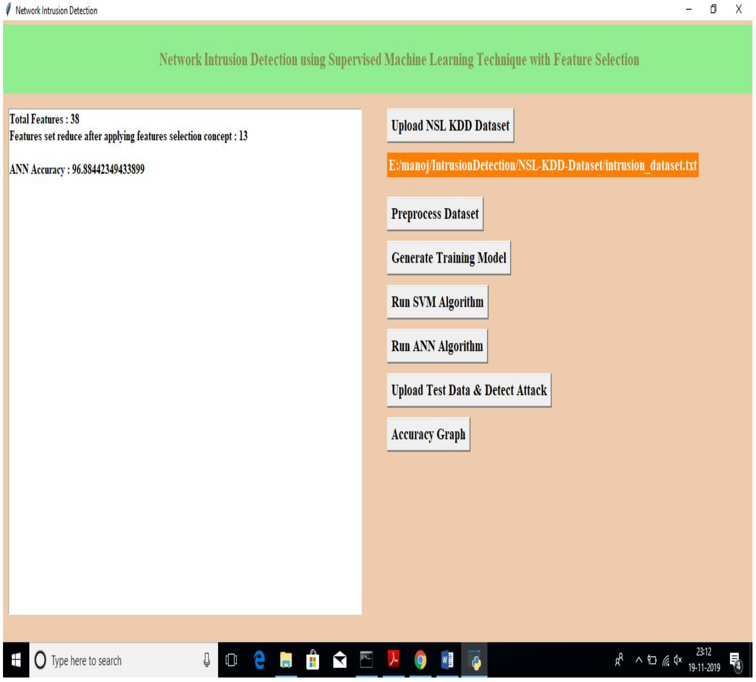
In above screen we can see dataset contains total 1244 records and 995 used for training and 249 used for testing. Now click on 'Run SVM Algorithm' to generate SVM model and calculate its model accuracy


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search
23:10
19-11-2019

E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_dataset.txt

```
Total Features : 38
Features set reduce after applying features selection concept : 0

Prediction Results

SVM Accuracy, Classification Report & Confusion Matrix

Accuracy : 84.73895582329317
```

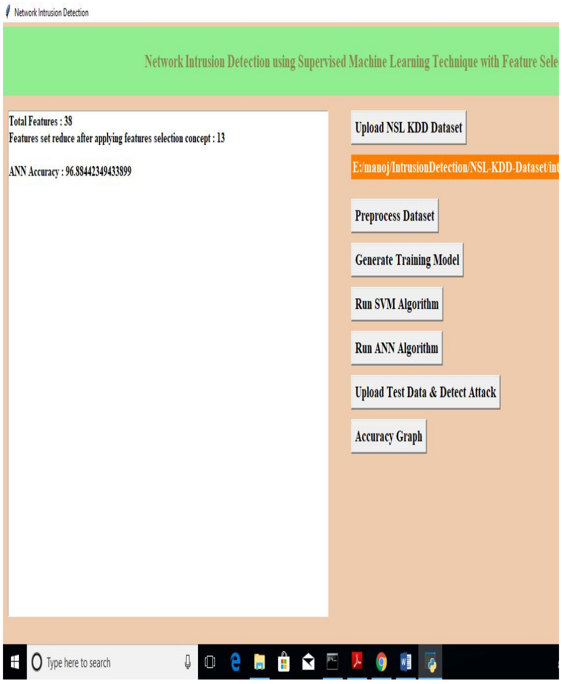
In above screen we can see with SVM we got 84.73% accuracy, now click on 'Run ANN Algorithm' to calculate ANN accuracy


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search
23:12
19-11-2019

E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_dataset.txt

```
Total Features : 38
Features set reduce after applying features selection concept : 13

ANN Accuracy : 96.88442349433899
```

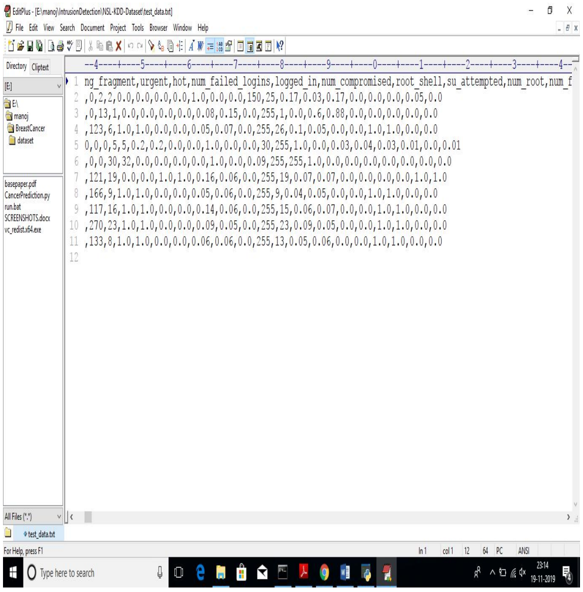
In above screen we can see with SVM we got 84.73% accuracy, now click on ‘Run ANN Algorithm’ to calculate ANN accuracy


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Sele
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search

E:/manoj/IntrusionDetection/NSL-KDD-Dataset/in

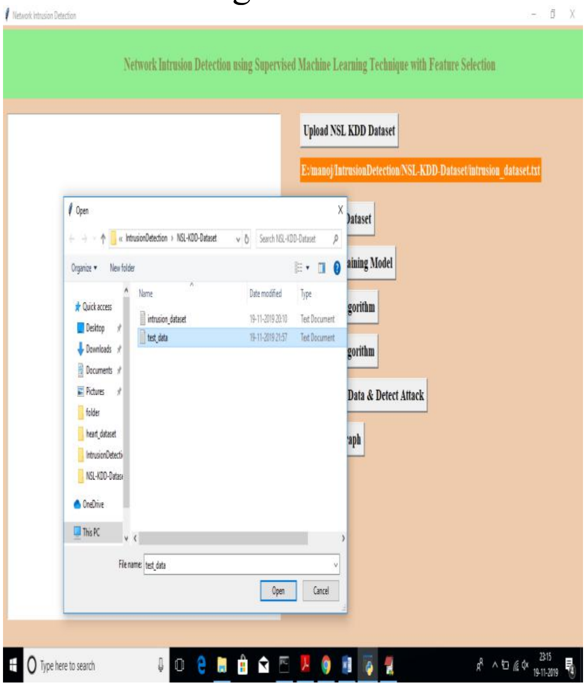
```
Total Features : 38
Features set reduce after applying features selection concept : 13

ANN Accuracy : 96.88442349433899
```

In above screen we got 96.88% accuracy, now we will click on ‘Upload Test Data & Detect Attack’ button to upload test data and to predict whether test data is normal or contains attack. All test data has no class either 0 or 1 and application will predict and give us result. See below some records from test data



```
ng_fragment,urgent,hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_f
,0,2,2,0.0,0.0,0.0,0.0,1.0,0.0,0.0,150,25,0.17,0.03,0.17,0.0,0.0,0.0,0.05,0.0
,0,13,1,0.0,0.0,0.0,0.0,0.08,0.15,0.0,255,1,0.0,0.6,0.88,0.0,0.0,0.0,0.0,0.0
,123,6,1.0,1.0,0.0,0.0,0.05,0.07,0.0,255,26,0.1,0.05,0.0,0.0,1.0,1.0,0.0,0.0
0,0,0,5,5,0.2,0.2,0.0,0.0,1.0,0.0,0.0,30,255,1.0,0.0,0.03,0.04,0.03,0.01,0.0,0.01
,0,0,30,32,0.0,0.0,0.0,0.0,1.0,0.0,0.09,255,255,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,121,19,0.0,0.0,1.0,1.0,0.16,0.06,0.0,255,19,0.07,0.07,0.0,0.0,0.0,0.0,1.0,1.0
,166,9,1.0,1.0,0.0,0.0,0.05,0.06,0.0,255,9,0.04,0.05,0.0,0.0,1.0,1.0,0.0,0.0
,117,16,1.0,1.0,0.0,0.0,0.14,0.06,0.0,255,15,0.06,0.07,0.0,0.0,1.0,1.0,0.0,0.0
,270,23,1.0,1.0,0.0,0.0,0.09,0.05,0.0,255,23,0.09,0.05,0.0,0.0,1.0,1.0,0.0,0.0
,133,8,1.0,1.0,0.0,0.0,0.06,0.06,0.0,255,13,0.05,0.06,0.0,0.0,1.0,1.0,0.0,0.0
```

In above test data we don’t have either ‘0’ or ‘1’ and application will detect and give us result

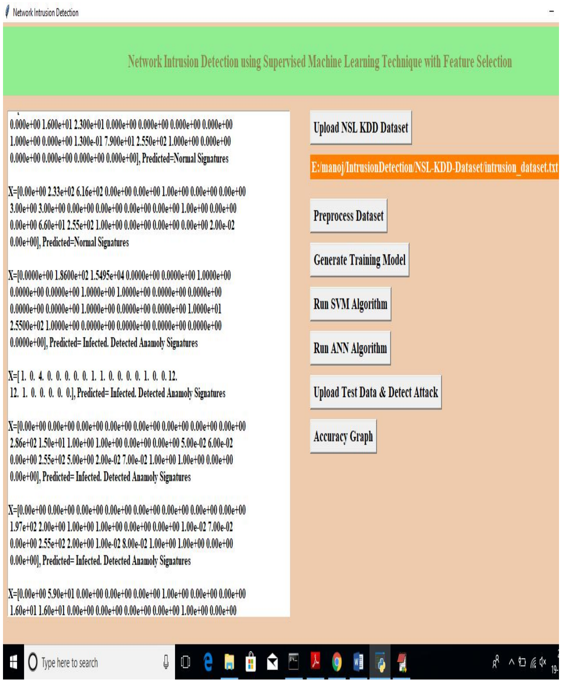


E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_dataset.txt

Name	Date modified	Type
intrusion_dataset	19-11-2019 20:10	Text Document
test_data	19-11-2019 23:57	Text Document

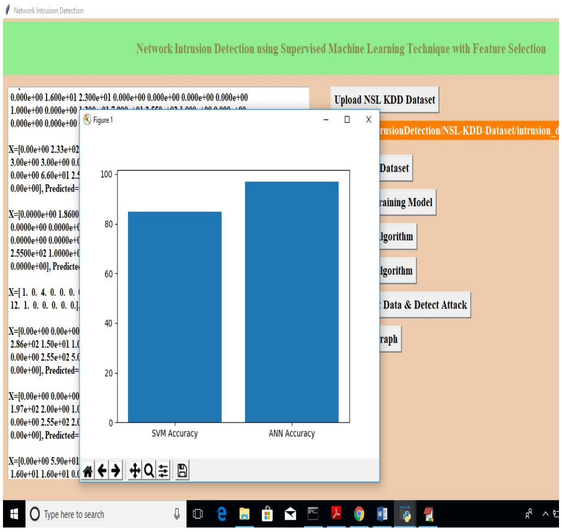
File name: test_data

In above screen I am uploading ‘test_data’ file which contains test record, after prediction will get below results


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
E:/manoj/IntrusionDetection/NSL-KDD-Dataset/intrusion_dataset.txt
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph
Type here to search

```
0.000e+00 1.600e+01 2.300e+01 0.000e+00 0.000e+00 0.000e+00 0.000e+00
1.000e+00 0.000e+00 1.300e-01 7.900e+01 2.550e+02 1.000e+00 0.000e+00
0.000e+00 0.000e+00 0.000e+00 0.000e+00], Predicted=Normal Signatures

X=[0.00e+00 2.33e+02 6.16e+02 0.00e+00 0.00e+00 1.00e+00 0.00e+00 0.00e+00
 3.00e+00 3.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 1.00e+00 0.00e+00
 0.00e+00 6.60e+01 2.55e+02 1.00e+00 0.00e+00 0.00e+00 0.00e+00 2.00e-02
 0.00e+00], Predicted=Normal Signatures

X=[0.0000e+00 1.8600e+02 1.5495e+04 0.0000e+00 0.0000e+00 1.0000e+00
 0.0000e+00 0.0000e+00 1.0000e+00 1.0000e+00 0.0000e+00 0.0000e+00
 0.0000e+00 0.0000e+00 1.0000e+00 0.0000e+00 0.0000e+00 1.0000e+01
 2.5500e+02 1.0000e+00 0.0000e+00 0.0000e+00 0.0000e+00 0.0000e+00
 0.0000e+00], Predicted= Infected. Detected Anamoly Signatures

X=[ 1. 0. 4. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 1. 0. 0. 12.
 12. 1. 0. 0. 0. 0. 0.], Predicted= Infected. Detected Anamoly Signatures

X=[0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00
 2.86e+02 1.50e+01 1.00e+00 1.00e+00 0.00e+00 0.00e+00 5.00e-02 6.00e-02
 0.00e+00 2.55e+02 5.00e+00 2.00e-02 7.00e-02 1.00e+00 1.00e+00 0.00e+00
 0.00e+00], Predicted= Infected. Detected Anamoly Signatures

X=[0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00
 1.97e+02 2.00e+00 1.00e+00 1.00e+00 0.00e+00 0.00e+00 1.00e-02 7.00e-02
 0.00e+00 2.55e+02 2.00e+00 1.00e-02 8.00e-02 1.00e+00 1.00e+00 0.00e+00
 0.00e+00], Predicted= Infected. Detected Anamoly Signatures

X=[0.00e+00 5.90e+01 0.00e+00 0.00e+00 0.00e+00 1.00e+00 0.00e+00 0.00e+00
 1.60e+01 1.60e+01 0.00e+00 0.00e+00 0.00e+00 0.00e+00 1.00e+00 0.00e+00
```

In above screen for each test data we got predicted results as 'Normal Signatures' or 'infected' record for each test record. Now click on 'Accuracy Graph' button to see SVM and ANN accuracy comparison in graph format


Network Intrusion Detection
Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Figure 1
100
80
60
40
20
0
SVM Accuracy
ANN Accuracy
Type here to search

From above graph we can see ANN got better accuracy compare to SVM, in above graph x-axis contains algorithm name and y-axis represents accuracy of that algorithms

CONCLUSION

we have presented different machine learning models using different machine learning algorithms and different feature selection methods to find a best model. The analysis of the result shows that the model built using ANN and wrapper feature selection outperformed all other models in classifying network traffic correctly with detection rate of 94.02%. We believe that these findings will contribute to research further in the domain of building a detection system that can detect known attacks as well as novel attacks. The intrusion detection system exist today can only detect known attacks. Detecting new attacks or zero day attack still remains a research topic due to the high false positive rate of the existing systems

REFERENCES

[1] H. Song, M. J. Lynch, and J. K. Cochran, "A macro-social exploratory analysis of the rate of interstate cyber-victimization," American Journal of Criminal Justice, vol. 41, no. 3, pp. 583–601, 2016.

[2] P. Alaei and F. Noorbehbahani, "Incremental anomaly-based intrusion detection system using limited labeled data," in Web Research (ICWR), 2017 3th

International Conference on, 2017, pp. 178–184.

[3] M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, “Modeling and implementation approach to evaluate the intrusion detection system,” in International Conference on Networked Systems, 2015, pp. 513–517.

[4] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, “Toward credible evaluation of anomaly-based intrusion-detection methods,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 5, pp. 516–524, 2010.

[5] A. S. Ashoor and S. Gore, “Importance of intrusion detection system (IDS),” International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1–4, 2011.

[6] M. Zamani and M. Movahedi, “Machine learning techniques for intrusion detection,” arXiv preprint arXiv:1312.2177, 2013.

[7] N. Chakraborty, “Intrusion detection system and intrusion prevention system: A comparative study,” International Journal of Computing and Business Research (IJCBR) ISSN (Online), pp. 2229–6166, 2013.

[8] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, “Anomaly-based network intrusion detection: Techniques, systems and challenges,” computers & security, vol. 28, no. 1–2, pp. 18–28, 2009.

[9] M. C. Belavagi and B. Muniyal, “Performance evaluation of supervised machine learning algorithms for intrusion detection,” Procedia Computer Science, vol. 89, pp. 117–123, 2016

. [10] J. Zheng, F. Shen, H. Fan, and J. Zhao, “An online incremental learning support vector machine for large-scale data,” Neural Computing and Applications, vol. 22, no. 5, pp. 1023–1035, 2013.

[11] F. Gharibian and A. A. Ghorbani, “Comparative study of supervised machine learning technique