APPLICATION OF MACHINE LEARNING IN CYBER SECURITY

BERNADINE
MANISHA S SONEJA
S SHIVARAJ
VAIBHAV GARG

PROF RESMA K S
Batch No. - 20

March 27, 2019

Problem Statement / Definition

MACHINE LEARNING FRAMEWORK FOR INTRUSION DETECTION SYSTEM

1. Domain: MACHINE LEARNING AND CYBER SECURITY
2. What: A Machine Learning framework for IDS
3. How: Using Various ML Algorithms
4. Why: To Detect if an attack is truly malicious or not and thereby reduce the false errors.

Motivation of the Work

- Internet is a house of huge amount of data thereby its security should be of primary concern. There are many softwares such as the SIEM , SoC that are responsible for governing the security of a network but these softwares arent very effiecient because they cannot handle false alerts thereby making it necessary to develop a mechanism that detects if the alert is actually malicious or not.
- Machine Learning can be used to implement an Intrusion Detection System to detect if an alert is truly malicious or not.

Literature Survey

[1] there are two basic ways to detect intrusions:

- **SIGNATURE BASED:** Compares the signature of an incoming packet with the database of existing signatures. If the signature matches the signatures in the database it is marked malicious. It is very effective in detecting known type of attacks, but is unable to detect new attacks.
- **ANOMALY BASED:** This is used to model a behaviour and if it deviates too much from normal it is marked malicious. Its a statistical based approach that is used to define normal behaviour. It works well will new attacks. But even if a new attack is just different from the normal it is marked malicious.

Literature Survey Continued

SUPPORT VECTOR MACHINE

It's a supervised machine learning algorithm. It provides a powerful method for classification. [2] , the basic approach in svm :

- The main goal seems to maximize the hyper plane. Hyper plane are the best separate between two classes and can Find by measuring the hyper plane margin and Find maximal data. Margin is the distance between the hyper plane with the closest data from every class. The closest data is nothing but the support vector

  [3]:

- In purposed framework NSL-KDD dataset is ranked using IGR and later feature subset selection is done using K-mean algorithm.

Literature Survey Continued

ADVANTAGES AND LIMITATIONS OF SVM:

Advantages:

- Good generalization nature, Ability to overcome curse of dimensionality
- Speed , capibility to work with real time environment.
- SVMs also have the ability to update the training pattern dynamically
- Its the best to classify abnormal behaviour.

Limitations:

- It treats every feature of data equally
- It requires labelled data which may not be available all the time
- Training of SVM is time-consuming for IDS domain and requires large dataset storage

Literature Survey Continued

GENETIC ALGORITHMS [4]:

- Based on identification of complex anomalous behaviours mainly focused on TCP/IP network protocol.
- The basic procedure:
    - Selection of chromosomes, represent the solutions
    - Different position of each chromosome are encoded as bits also called genes
    - Fitness function / Evaluation Function: Used to evaluate the goodness of each chromosome.
    - Two basic operators: Crossover and Mutation
    - Based on rule of "Survival Of the Fittest - Darwin Theory"
- It performs a match between the current n/w and rules in IDS such as src/dest IP address indicating probability of intrusion
  Action: Report to Admin, Stop the connection

Literature Survey Continued

HYBRID CLASSIFIERS: [5]:

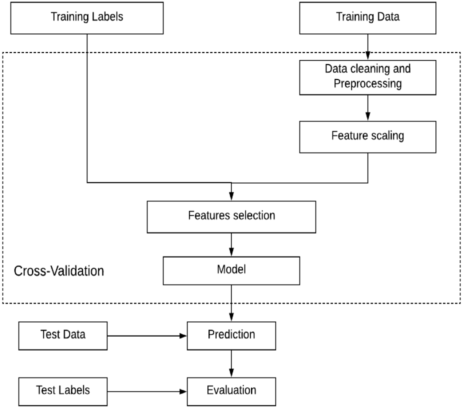
- Collects every single incoming network requests
- The basic procedure:
  - A decision tree algorithm is used to train the Misuse Detection model from the available normal and attack training data
  - Info Gain is calculated for all attributes to understand the contribution of the attributes towards the classification.
  - After training , its checked against the DT Model to classify if it's an attack or not, if there is no response it is sent to Naive bayes classifier - anomaly detection.

  Advantages:
- Least Number of false positive as it has to go through DT and then Naive Bayes model
- Only the important attributes are considered, according to info gain leading to attribute reduction
- Limitation: Increased Time Complexity for IDS

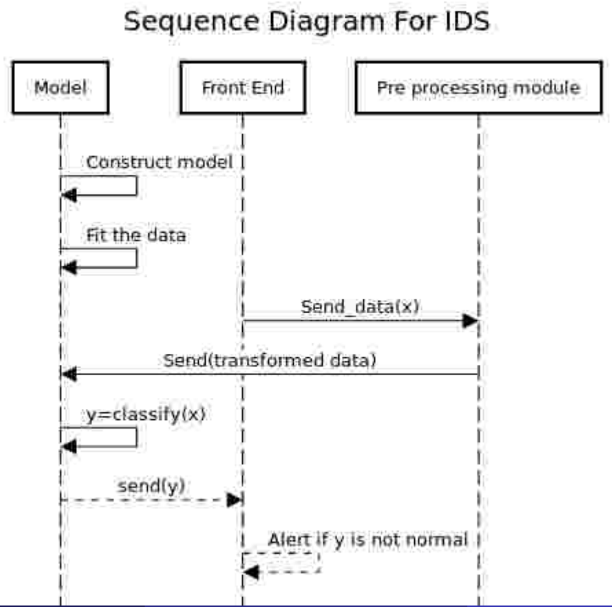
Methodology

- Cleaning Data: NSL KDD Dataset
    - It's an improvement over kdd data set as it doesn't have any redundant records.
    - Data Cleaning is not required
    - Categorical attributes need to be one-hot encoded
- Scaling the Dataset
- Feature selection
- Applying C4.5 Algorithm
- Naive Bayes
- Prediction and Evaluation on test Data

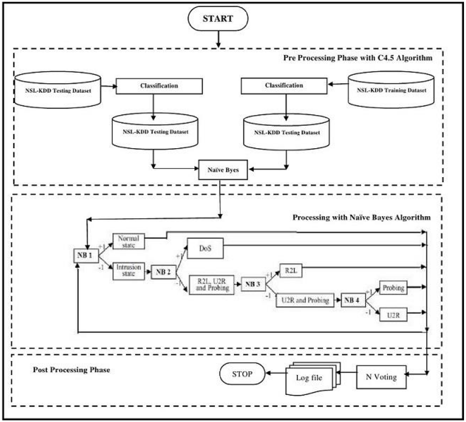
High Level Design



Sequence Diagram



Detailed Design



Hardware And Software Requirements

Hardware Requirements:

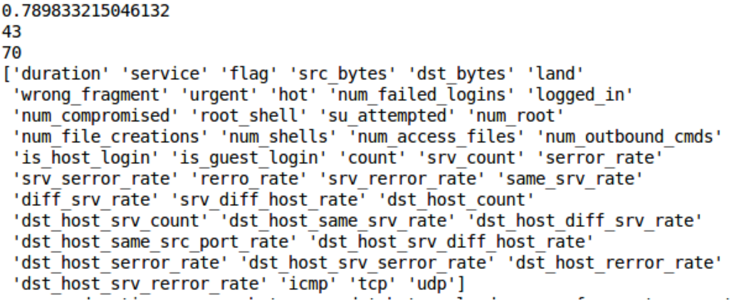
- Processor: Intel Core i5 or higher
- Main Memory-8.0GB RAM
- HardDisk - 1TB
- Monitor- LCD

Software Requirements:

- Operating System Platform- Linux
- Language: Python (Backend), Front End -HTML5/CSS

Results and Snapshots

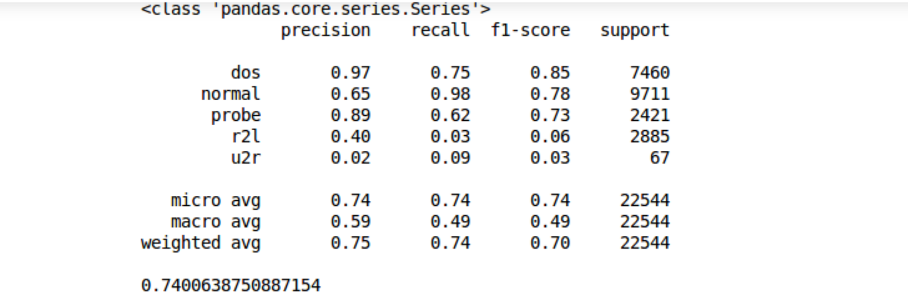
After Decision Tree Algorithm



```
0.789833215046132
43
70
['duration' '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' 'rerro_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' 'icmp' 'tcp' 'udp']
```

Results and Snapshots Continued

After Naive Bayes Algorithm



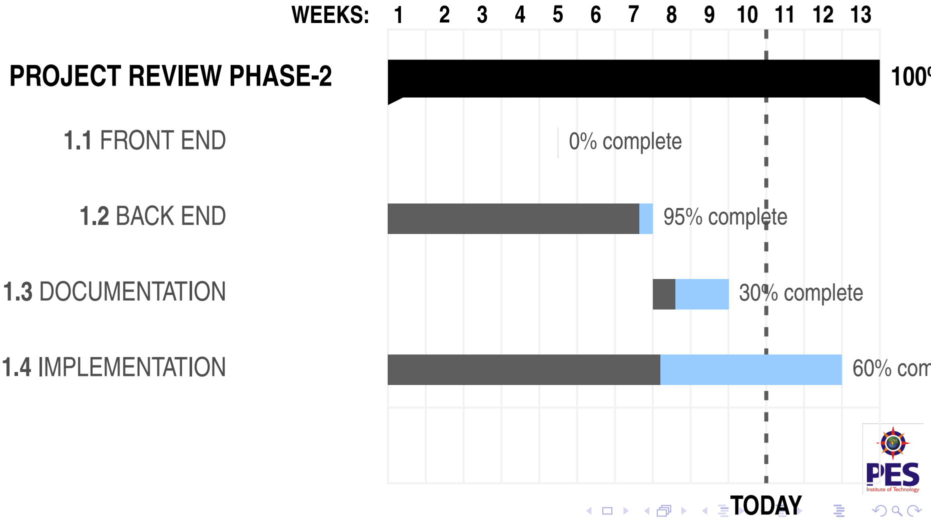
```
<class 'pandas.core.series.Series'>
              precision    recall  f1-score   support

         dos       0.97      0.75      0.85      7460
      normal       0.65      0.98      0.78      9711
       probe       0.89      0.62      0.73      2421
         r2l       0.40      0.03      0.06      2885
         u2r       0.02      0.09      0.03        67

   micro avg       0.74      0.74      0.74     22544
   macro avg       0.59      0.49      0.49     22544
weighted avg       0.75      0.74      0.70     22544

0.7400638750887154
```

Time line of completion of project from 3rd January 2019-20th March 2019(Gantt Charts).



References

[1] Intrusion Detection System using AI and Machine Learning Algorithm(Dec-2017)
Syam Akhil Repalle, Venkata Ratnam Kolluru
IRJET

[2] Study on Implementation of Machine Learning Methods Combination for Improving Attacks Detection Accuracy on Intrusion Detection System (IDS)
Bisyron Wahyudi Masduki, Kalamullah Ramli,Ferry Astika Saputra, Dedy Sugiarto
IEEE-2015-7374895

[3] Intrusion Detection System using Support Vector Machine
Jayshree Jha ,Leena Ragha, Ph.D
International Journal of Applied Information Systems (IJAIS) ISSN : 2249-0868

[4] Using Genetic Algorithm for Network Intrusion Detection
Wei Li
Mississippi State University, Mississippi State, MS 39762

References Continued..

[5] An Intrusion Detection System, (IDS) with Machine Learning (ML) Model Combining Hybrid Classifiers

Arjunwadkar Narayan M,Thaksen J. Parvat

Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 (2015)

[6] A User Centric Machine Learning Framework for Cyber Secrity Operations Center(2017)

Charles Feng ,Shuning Wu

IEEE 08004902.

[7] Network Intrusion Detection Using Machine Learning

Md Nasimuzzaman Chowdhury and Ken Ferens, Mike Ferens

ISBN: 1-60132-445-6, CSREA Press

The End