NAME - ARNAV KAUSHIK SHARMA ID - 201501236

NETWORK INTRUSION DETECTION SYSTEM

Problem Description

A network intrusion is any unauthorized activity on a computer network. It absorbs network resources intended for other uses, and nearly always threatens the security of the network and/or its data. Properly designing and deploying a NIDS will help block the intruders. In this project, the NIDS is developed in two parts – First, to differentiate between normal connection and an intrusion and Second, to find the category of various types of attacks.

Motivation

Intrusion Detection is one of the major concerns in the task of network administration and security. There is a need to safeguard the networks from known vulnerabilities and at the same time take steps to detect new and unseen, but possible, system abuses by developing more reliable and efficient IDS. This is the need of the present and it poses a challenging problem that needs to be solved more efficiently.

Summary of Research Papers

There are 3 types of IDS – Host based, Network based & Application based. Network based IDS collects information from the network itself rather than from each separate host. NIDS audits the network attacks while packets moving across the network. The network sensors come equipped with attack signatures that are rules on what will constitute an attack and most network-based systems allow advanced users to define their own signatures. Attack on the sensor is based on signature and they are from the previous attacks and the operation of the monitors will be transparent to the users and this is also significant.

Advantages of Network based Intrusion Detection Systems: -

- Lower Cost of Ownership
- Easier to deploy
- Detect network based attacks
- Retaining evidence
- Real Time detection and quick response
- Detection of failed attacks

Functions of NIDS: -

- **Data Collection and Pre-processing** – Collect the details of data packets in a file to pre-process and alter them
- **Feature Selection** – Select the important features to evaluate the robustness of the network security
- **Analysis** – One way is that incoming traffic is checked against predefined signature or pattern. Another way is anomaly based, where the system behaviour is studied and mathematical models are employed to it.
- **Action** – Alert the administrator about possible intrusion or actively prevent it.

Various methods have been proposed for building a Network Intrusion Detection System, which are based on: -

- SVM (Support Vector Machine)
- Naïve Bayes Classifier
- Decision Trees
- Random Forests
- kNN (k Nearest Neighbours) Algorithm
- KPDS (kNN Partial Distance Search) Algorithm
- IKPDS Algorithm

Variance based Feature Selection (FS) can be used to evaluate the quality and predictive power of the features and select those whose variance is higher than the threshold, since they are best for use.

Dataset

KDD Cup 1999 Dataset was used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99, The Fifth International Conference on Knowledge Discovery and Data Mining.

It has 41 attributes which describe the various conditions of a network connection. Each connection is initially classified as either 'normal' or one of the other 22 attacks.

All the 23 types of attacks are classified into 5 major classes of network intrusions: -

1. **normal** – normal (normal network connection – NOT AN INTRUSION)

2. **dos** – smurf, neptune, back, teardrop, pod, land

3. **probe** – satan, ipsweep, portsweep, nmap

4. **u2r** – buffer_overflow, rootkit, loadmodule, perl

5. **r2l** – spy, phf, multihop, ftp_write, imap, warezmaster, guess_passwd, warezclient

Classification of connections

All 23 types of connections are classified into 5 major classes: -

- **DoS (Denial of Service)** – Computing resources are too busy to handle legit requests.
- **Probe Attacks** – Gather info about network of computers for circumventing its security controls.
- **R2L (Remote to User) Attacks** – Gain local access as a user to a machine remotely.
- **U2R (User to Root) Attacks** – Attacker starts with user access and then exploits some vulnerability to gain root access.
- **Normal Connections**

Algorithms and Results

1. **Support Vector Machine (SVM)** is used to build the class model with 3 parameters to set – '**kernel**', '**class_weight**' & '**max_iter**'.
    a. The values of the 3 parameters selected are kernel = linear, class_weight = balanced and max_iter = 10^8.

    b. 3 different measures are calculated for finding out the accuracy of the model – '**precision**', '**recall**', '**f1-score**'. The average values of the measures are: -

        i) **Precision** = 0.83
        ii) **Recall** = 0.88
        iii) **f1-score** = 0.85

2. Then, **Decision Tree** is built.
    i) **Entropy** = 0.9956
    ii) **Gini Index** = 0.946
    iii) **Overall Accuracy** = 0.7809

3. Then, **Naïve Bayes** Classification is applied. The measures calculated for finding out the accuracy of the model are '**precision**', '**recall**', '**f1-score**'. The average values of the measures are: -
    i) **Precision** = 0.81
    ii) **Recall** = 0.75
    iii) **f1-score** = 0.76

All these processes were run in 2 stages: -

1st stage – Classifying various network connections into 5 major categories of intrusions – 'DoS', 'Probe', 'R2L', 'U2R', 'Normal', i.e., identifying the different types of intrusions.

2nd stage – Classifying various network connections into 2 major categories – 'normal' and 'attack'.

4. Then, **kNN Algorithm** is applied. At k = 3,
    i) At k = 3,
    Overall Accuracy = 0.9698
    ii) At k = 5,
    Overall Accuracy = 0.9451

5. Then, **Random Forest** is applied.
    i) **Overall Accuracy** = 0.9978
    ii) **Precision** = 0.9458
    iii) **Recall** = 0.8364
    iv) **f1-score** = 0.8897

Conclusion

1. Random Forest gave the highest accuracy at 0.9978 and its f1-score was also highest at 0.8897.
2. kNN gave the second highest accuracy at 0.9698, but took a lot of time to predict the class of attack.
3. SVM with a linear kernel acted almost as linear regression, and gave almost the same precision as Naïve Bayes, at a little more than 0.81, mostly because attacks in the R2L class were very infrequent compared to Dos and Probe.
4. The accuracy in Decision Tree was the lowest at 0.78, probably because of over-fitting. This problem of over-fitting was solved by Random Forest.
5. The computation time of kNN can be reduced by the application of KPDS and IKPDS.

References

1. [Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets]

https://pdfs.semanticscholar.org/1d6e/a73b6e08ed9913d3aad924f7d7ced4477589.pdf

2. [A Detailed Analysis of the KDD CUP 99 Data Set]

https://www.ee.ryerson.ca/~bagheri/papers/cisda.pdf

3. [INTRUSION DETECTION SYSTEM – A STUDY**]**

https://airccse.org/journal/ijsptm/papers/4115ijsptm04.pdf

4. [Fast kNN Classifiers for Network Intrusion Detection System]

https://ijact.org/volume2issue4/IJ0240025.pdf

5. [The 1998 DARPA Intrusion Detection Evaluation]

https://www.ll.mit.edu/ideval/files/Evaluating_IDs_DARPA_1998.pdf