International Conference on Research Advances in Integrated Navigation Systems (RAINS - 2016),
April 06-07, 2016, R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India

Network Intrusion Detection System Using various data mining techniques

DikshantGupta[1] SuhaniSinghal[2] Shamita Malik[3] Archana Singh[4]
Amity School of Engineering and Technology, Amity University, Uttar Pradesh
[1]dikshantgupta99@gmail.com
[2]s08s03s94@gmail.com
[3]mailtoshamita07@gmail.com
[4]archana.elina@gmail.com

Abstract: There are many risk of network attacks in the Internet environment. Nowdays, Security on the internet is a vital issue and therefore, the intrusion detection is one of the major research problem for business and personal networks which resist external attacks. A Network Intrusion Detection System (NIDS) is a software application that monitors the network or system activities for malicious activities and unauthorized access to devices. The goal of designing NIDS is to protect the data's confidentiality and integrity. Our project focuses on these issues with the help of Data Mining. This research paper includes the implementation of different data mining algorithms including Linear regression and K-Means Clustering to automatically generate the rules for classify network activities. A comparative analysis of these techniques to detect intrusions has also been made. To learn the patterns of the attacks, NSL-KDD dataset has been used.

KEYWORDS: Network Intrusion Detection System (IDS), Data Mining Techniques.

I. INTRODUCTION

An Intrusion Detection System (IDS) is a software application that monitor networks or system activities for all the malicious activities and unauthorized access to devices. IDS come in a variety of "flavors" and it aims at detecting suspicious traffic in many ways. There are networks based (NIDS) and host based intrusion detection system (HIDS). There are some IDS devices that detect attacks based on comparing traffic patterns against a baseline and then looks for anomalies. There are some IDS that simply monitor and alert and there are some IDS which performs an action or actions in response to a detected threat. We will include each of these briefly. Network Intrusion Detection Systems are placed at strategic point or points within the network to monitor traffic to and from all the devices on the network. Ideally you would inspect all the inbound and outbound traffic; however doing so may create bottleneck that would impair the overall speed of the network.

II. DATA SET COLLECTION AND PRE-PROCESSING

A. Data Set Collection

To verify the efficiency and feasibility of the proposed IDS system, we have used NSL-KDD dataset.

It is a brand new version of KDDcup99 dataset. NSL-KDD dataset have some advantages over KDDcup99 dataset. It has solved some of the inherent problems of the KDDcup99, which is considered as the standard benchmark meant for intrusion detection assessment. The training dataset of NSL-KDD is similar to KDDcup99 which consist of approximately 4,900,000 single connection vectors each of which

978-1-4673-8819-8/16/$31.00 ©2016 IEEE

contain 41 features and are labeled as either normal or attack type, with exactly one specific attack type.

Due to the following reasons, NSL-KDD hasturn out to be more popular dataset than KDD cup 99 dataset for intrusion detection purpose[1].

1. Redundant records from the training set are eliminated.

2. Duplicate records from the test sets are removed to enhance the intrusion detection performance.

3. Use of NSL-KDD dataset for classification gives aprecise evaluation of different learning techniques.

4. It is affordable to use NSL-KDD dataset for the experiment purpose and also it consists of reasonable numbers of instances both in the training set as well as in the testing set.

B. Data Set Pre-Processing

Pre-processing is done of original NSL-KDD dataset which is necessary to make it as suitable input for our algothrims. Data set pre-processing is achieved by applying:

i. Data set transformation
ii. Data set normalization

i) Data set transformation: The training dataset of NSL-KDD which contains approximately 4,900,000 single connection instances. Each of the connection instance contain 42 features including attacks or normal. From these labeled connection instances, we need to convert the nominal features to numeric values so that it suitable input for classification using data mining techniques[2]. For this transformation, we will be using table 2. Also, we have to assign a numeric value to the last feature within the connection instance which is a target class. For doing this, we have assigned a target class zero for normal connection and one for any deviation from that (i.e. if that is an attack) as per transformation table 2.In this step, some useless data is filtered and then modified. For example, some text items need to be converted into numeric values. Each and every instance in the dataset have 42 features or attributes including target class shown in Table 1.

I. TABLE
FEATURES OF NSD- KDD CUP'99 DATASET

Sr.No	Feature Name
1	Duration
2	Protocol_type
3	Service
4	Flag
5	Src_bytes
6	Dst_bytes
7	Land
8	Wrong_fragment
9	Urgent
10	Hot
11	Num_failed_logins
12	Logged_in
13	Num_compromised
14	Root_shell
15	Su_attempted
16	Num_root
17	Num_file_creations
18	Num_shells
19	Num_acess_files
20	Num_outboun_cms
21	Is_host_login
22	Is_guest_login
23	Count
24	Srv_count
25	Serror_rate
26	Srv_serror_rate
27	Rerror_rate
28	Srv_rerror_rate
29	Same_srv_rate
30	Diff_srv_rate
31	Srv_iff_host_rate
32	Dst_host_count
33	Dst_host_srv_count
34	Dst_host_same_srv_rate
35	Dst_host_diff_srv_rate
36	Dst_host_same_src_port_rate
37	Dst_host_srv_diff_host_rate
38	Dst_host_serror_rate
39	Dst_host_srv_serror_rate
40	Dst_host_rerror_rate
41	Dst_host_srv_rerror_rate
42	Normal or Attack

An example of original NSL-KDD data set record is shown in figure 1.

```
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 udp other SF 146 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 13 1 0 0 0 0 0.08
0.15 0 255 1 0 0.6 0.88 0 0 0 0 0 normal
```

Figure 1: The Original NSL KDD cup'99 dataset

II. TABLE
TRANSFORMATION TABLE

Type	Feature Name	Numeric Value
Attack or Normal	Normal	0
Attack or Normal	Attack	1
Protocol_type	TCP	2
	UDP	3
	ICMP	4
Flag	OTH	5
	REJ	6
	RSTO	7
	RSTOS0	8
	RSTR	9
	S0	10
	S1	11
	S2	12
	S3	13
	SF	14
	SH	15
Service	All Services	16 to 81

There are numerous nominal values like http, tcp, SF in the dataset. Hence,we have to convert these nominal values to numeric values in advance. For example, service type like "tcp" is mapped to 2,"udp" is mapped to 3,and "icmp" is mapped to 4 and then we will follow Table 2 to transform these nominal values of the dataset features into the numeric values[3]. After transformation, the original NSL- KDD cup‟99 dataset will become as shown in Figure 2.

```
0,2,32,14,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0,0,0,0,1,0,
0,150,25,0.17,0.03,0.17,0,0,0,0.05,0,0

0,3,16,14,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,13,1,0,0,0,0,0.
08,0.15,0,255,1,0,0.6,0.88,0,0,0,0,0,0
```

Figure 2: Original NSL- KDD cup'99 dataset after transformation

ii) Data set normalization: Dataset normalization which is considered necessary to enhance the performance of intrusion detection system when datasets are too large.

Mean normalization is used that makes the values of each feature in the data have zero-mean and unit variance Normalization refers as the creation of shifted and the scaled versions of statistics, where the intention is that the normalized value allows the comparison of the corresponding normalized values for different datasets in a way that eliminates all the effects of certain gross influence, as in an anomaly time series. Some types of normalization involve only rescaling so that to arrive at the values relative to some size variable. In terms oflevels of measurement, such ratios just make sense for *ratio* measurements (where ratios of measurement are meaningful), not *interval* measurements (where only distances are meaningful, but not ratios)[4]. Assume that there are *n* rows with only seven variables (columns), A, B, C, D, E, F and G, in the data. We use a variable E as an example in the calculations below. The remaining variables in the row are normalized in the same way only.

The normalized value of e_i for variable E in the i^{th} row is calculated as:

$$Normalized\ (e_i) = \frac{e_i}{\frac{1}{p}\ \sum_{j=1}^{p} e_j}$$

Where,

p = it is the number of records that is used to calculate the mean.

III. Application of Algorithms

i. Linear Regression:

∑ Study of linear relationship between an output variable and one or more input features.

∑ For a linear model, our hypothesis of the form,
$h_\theta(x) = \theta_0 + \theta_1 x$ or $h_\theta(x) = \theta^T x$.

∑ We have to adjust θ_0 and θ_1 so that h_θ(x) is close to y.

For that we define an error function, $J(\theta_0, \theta_1) = (1 / 2m) * \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$.

∑ To minimize the error, we use Gradient Descent Algorithm,
Repeat until convergence $\{\theta_j = \theta_j - \alpha\frac{\partial}{\partial\theta_0} J(\theta_0, \theta_1)$ for j=0 and j=1}

(ii) K-Means Clustering

- It is an algorithm to categorize or to group all the objects based on the attributes/features into K number of groups.
- The grouping which is done by minimizing the sum of the square of distances between data and the corresponding cluster centroid.[5]
- Algorithm

```
Randomly initialize the K cluster centroids
Repeat {
        for i=1 to m
        c(i) := j that minimizes $||x(i) - \mu_j||^2$,
        for k=1 to K
        $\mu_k$ = average mean of the points assigned to cluster k.[6]
```

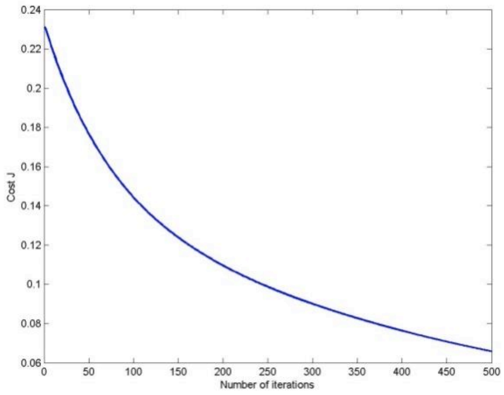
IV. Results

1. Linear Regression Results:

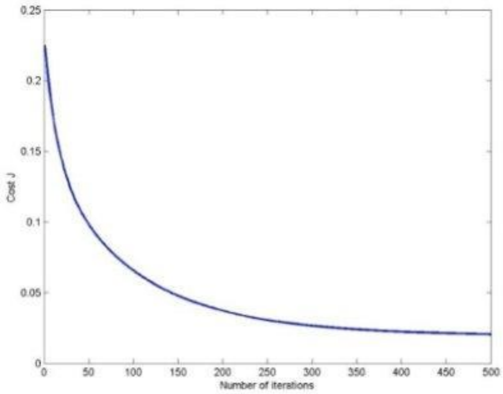
For different values of alpha, linear regression algorithm was trained to find the best result.

Alpha(α)	0.001	0.005	0.01	0.03	0.1	1
Accuracy (%)	79.20	80.14	71.47	70.98	69.65	65.5

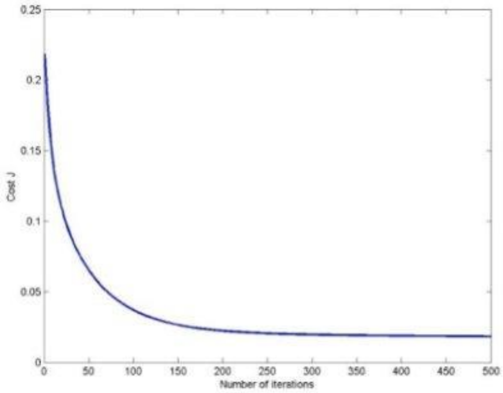
∑ Cost Variations (alpha = 0.001)



∑ Cost Variation (alpha = 0.005)



∑ Cost Variation (alpha = 0.01)



2. K-means Clustering Results:

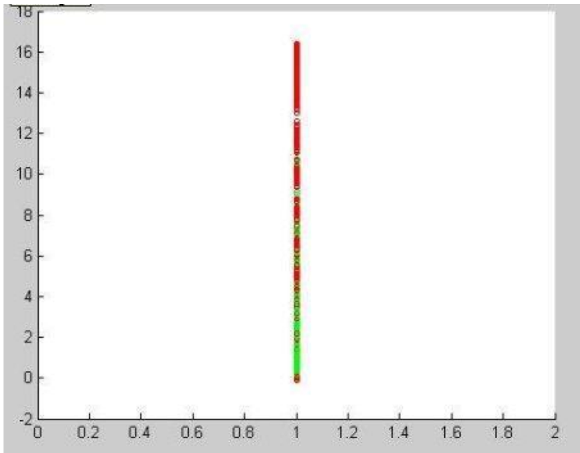
∑ We used two clusters, one to classify attacks and other for normal patterns.

∑ Random centroids were selected and there results were compared.

∑ Best centroid value was found to be X(i = 84252, :) and X(j = 17428, :)

∑ Accuracy was found 67.53%.

Cluster Plot



V. Conclusions

Algorithms based on Data Mining were implemented successfully showing different accuracies..NSL-KDD dataset was preprocessed using mean normalization method.Linear regression, surprisingly, proved to be very effective in detecting network attacks with a 80% accuracy. K-Means Clustering being a semi-supervised approach showed decent results with a 67.5 % accuracy.

VI. FUTURE WORKS

∑ Network Intrusion Detection can be improved using latest data mining techniques like deep neural network, dbscan etc.

∑ Here we have not used feature selection to select only relevant features. Using feature selection time performance can be improved as well as accuracy.

∑ Dimensionality Reduction using principal component analysis can be used to improve the time and visualization.

∑ The NSL-KDD dataset is very old and there are some bugs which can be tackled using a well derived dataset.

VII. References

1. Tavallaee, Mahbod, et al. "A detailed analysis of the KDD CUP 99 data set."Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009. 2009.
2. Bhavsar, Yogita B., and Kalyani C. Waghmare. "Intrusion detection system using data mining technique: Support vector machine." International Journal of Emerging Technology and Advanced Engineering 3.3 (2013): 581-586.
3. Siddiqui, Mohammad Khubeb, and Shams Naahid. "Analysis of KDD CUP 99 dataset using Clustering based Data Mining." International Journal of Database Theory and Application 6.5 (2013): 23-34.
4. Revathi, S., and A. Malathi."A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection."*International Journal of Engineering Research and Technology*.Vol. 2.No. 12 (December-2013).ESRSA Publications, 2013.
5. Solanki, Miss Meghana, and MrsVidyaDhamdhere. "Intrusion Detection System by using K-Means clustering, C 4.5, FNN, SVM classifier."

6. Dabas, Poonam, and Rashmi Chaudhary. "Survey of Network Intrusion Detection Using K-Mean Algorithm." *International Journal of Advanced Research in Computer Science and Software Engineering* 3.3 (2013): 507-511.