International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 08 | Nov-2015 www.irjet.net p-ISSN: 2395-0072

Detection of Network Intrusions with PCA and Probabilistic SOM

Palakollu Srinivasarao

M.Tech, Computer Networks and Information Security, MVGR College Of Engineering, AP, INDIA

Abstract–*In the present society, the use of internet has increased drastically and hence it leads to give valid information to intruders and attackers. Therefore in order to detect attacks one must use a firewall. But the aim of firewall is to detect violations according to a predefined rule-set and usually block possibly threatening incoming traffic. However, with the invention of different attack techniques, it is more difficult to identify anomalies from normal traffic. So to protect our network system from these threats, it becomes very important to build up a system that acts as a barrier between the network systems and the unessential security attacks. For monitoring and detecting the intrusions an intrusion detection system (IDS) were developed. But the expected performance and accuracy are not achieved by these systems. In this paper we propose a statistical approach i.e. Principle Component Analysis (PCA) and Probabilistic SOM. The PCA and FDR have been considered for feature selection and noise removal. The SOM aims to detect anomalies from normal traffic. In addition to this we classify the four types of attacks presented in the network dataset. The dataset used for training and testing purposes is KDD dataset.*

Key Words: *PCA, FDR, SOM, intrusion detection, feature selection.*

1. INTRODUCTION

In present days, working with internet, and as well as with network applications has increased drastically. On the other hand, the trend to online services has exposed valid information to intruders and attackers. The complexity of the newer attacks necessitates the use of elaborated techniques, such as pattern classification or artificial intelligence for successfully detecting an attack or just to differentiate among normal and anomalous traffic. Perimeter security involves IDS, IPS, and Firewalls etc. In [1] the authors proposes that, IDS, IPS and firewalls are the active systems. The main aim of these systems is to monitor the network in order to classify the traffic, as normal or abnormal behaviour. Most of these systems will execute some features from the monitored network and react according to predefined rules. This problem is looking like a classification problem, thus contributing to machine learning field. In [2] the author specifies that, security is the key step for network system. There are different soft computing methods have been expressed in the last decades but these methods will fail to detect attacks, which are vary from the examined patterns. In [3] the authors specifies that, PCA is useful technique for finding patterns in the data. Therefore Artificial Neural Networks (ANN) comes into existence, which can detect attack with a limited, nonlinear data sources. Intrusion detection is not a straight forward task so different detection approaches came into existence including the use of ANN such as neural networks. In real world environment intrusion detection poses different problems related to feature selection and classification. In this work we considered the network dataset such as NSL KDD dataset for experiment but this dataset contain huge amount of records so we will take only test set with some sub set and training sub set. In this process we will working with PCA in order to select a sub set of features from the training dataset. On the other hand, SOM enables abnormal connections from the normal connections. The section I specifies that introduction to our contribution, section II illustrates that, the intrusion detection and the dataset we considered for our work, section III specifies that proposed methods for anomaly detection and finally section IV represents our results with conclusions.

1.2 RELATED WORK

There are many researchers on network intrusion detection methods one of those [1], specifies that feature selection plays a major role in anomaly field and he proposed PCA as a feature selection method. In which he selected 15 features over the 41 features of kdd dataset. In [4] the author represents that, an intrusion detection based on machine learning approaches as 2 class classification. In [5]the author specifies that complete survey on SOM based IDS.In [6] the authors proposed a

novel approach of attribute reduction based on rough sets and ant colony optimization.

2. INTRUSION DETECTION SYSTEM AND DATASET

2.1 Intrusion Detection system

Intrusion detection systems are plays a major role in security as it tends to protect our network from unauthorized access. The most important intrusions includes Dos, DDos, man in the middle attack, masquerading etc. The IDS aims to detect this type of attacks with different attack techniques such as neural networks. Generally, there are two types of detection methods based on modelling.one is[7] misuse detection and second is anomaly detection. The misuse detection will detect anomalies with predefined rules so we can call it as virus scanners. Although misuse detection is effective for detecting known attacks, it cannot detect new attacks that were not before defiend.Anomaly detection aims to detect new attacks.

IDS can be classified as network based or host based IDS depending on target environment. Network based IDS aims to monitor the network behaviour by examining the content as well as the format of network traffic. Host based IDS will examining the host system information such as CPU time, system calls and command sequences.

2.2 KDD CUP'99 dataset description

The researchers who are looking in intrusion detection area will be looking forward to use a network dataset and hence the answer is KDD cup99 dataset. As we know, KDD cup'99 dataset is the standard dataset which is mostly used in anomaly detection methods. This dataset was provided by DARPA '98. The KDD dataset was generated by observing a military network composed of three machines working on three different os and traffic. This traffic has been captured by a sniffer in the format of TCP dump. The total capture time was seven weeks and there are some normal connections and some abnormal connections. The attacks can be broadly classified into four categories.[8]

2.2.1 Denial of service attack (Dos)

Dos is the attack, in which the attacker tries to send malicious packets such as tcp, udp, or icmp in order to fill up the memory or to make the computing resources busy to avoiding correct user to use a machine. Eg. Syn – flooding, Neptune, back, smurf.

2.2.2 U2R (User to Root)

In this attack type, the attacker may have local access to the target machine and hence tries to gain super user privileges. Eg. Buffer overflow

2.2.3 R2L (Remote to Local)

R2L is a special type of attack in this the attacker wants to access the target machine without having any privileges. Eg. ftp_write, guess_pwd.

2.2.4 Probing Attack

Probing attack is the one, in which the hacker aim to acquire the information about network of computers. Eg. nmap, ipsweep

The KDD dataset contains 41 features for each packet. Here we are presenting the description of each packet in Table.1.However a link is a structure of TCP packets starting and ending with some well-defined time intervals between which data flows from source IP address to destination IP address based on some well-defined protocols. The features can be classified into 4 categories.

- Basic Features: These Features can be derivative from the packet headers without inspecting the payload
- Content Features: The domain knowledge is used to access the payload of the original TCP packets. This includes features such as number of failed login attempts.
- Time based Traffic Features: Number of connections to the same host over 2 second interval
- Host based Traffic Features: These Features were designed to detect attack such as span intervals longer than two seconds

Table-1: KDD cup99 dataset features

No.	Features	No.	Features
1.	duration	22.	Is guest login
2.	protocol type	23.	count
3.	service	24.	Srv count
4.	flag	25.	Serror rate
5.	src bytes	26.	Srv serror rate
6.	dst bytes	27.	Rerror rate
7.	land	28.	Srv rerror rate
8.	wrong fragment	29.	Same srv rate
9.	urgent	30.	Diff srv rate
10.	hot	31.	Srv diff host rate
11.	num failed logins	32.	Dst host count
12.	Logged in	33.	Dst host srv count
13.	Num compromised	34.	Dst host same srv rate
14.	Root shell	35.	Dst host diff srv rate
15.	Su attempted	36.	Dst host same src port rate
16.	Num root	37.	Dst host srv diff host rate
17.	Num file creations	38.	Dst host serror rate
18.	Num shells	39.	Dst host srv serror rate
19.	Num access files	40.	Dst host rerror rate
20.	Num outbound cmds	41.	Dst host srv rerror rate
21.	Is host login		

Table-2: Basic characteristics of KDD '99 Intrusion Detection Datasets in terms of number of samples

Dataset	DoS	Probe	U2R	R2L	Normal
10% KDD	391458	4107	52	1126	97277
Corrected KDD	229853	4166	70	16347	60593
Whole KDD	3883370	41102	52	1126	972780

Table -3: Attack types with their corresponding categories

Category	Attack types
Probe	ipsweep, mscan, nmap, portsweep, saint, satan
DoS	apache, back, land, mailbomb, neptune, pod, processtable, smurf, teardrop, udpstorm
U2R	buffer_overflow, loadmodule, perl, rootkit, ps, sqlattack, xterm
R2L	ftp_write, guess_password, imap, multihop,

3. Proposed methods

In real world environment anomaly detection is not an easy task and which leads to several problems such as feature selection and classification. In this section we discuss the proposed anomaly detection methods.

3.1 Feature selection with PCA

In classification problems, feature selection [9] is the most crucial step as this process will remove the redundant or not [10]relevant features for the input The aim of removing irrelevant features is that, it increases the performance of the classifier accuracy. [11] The feature selection methods can be broadly classified into three types. They are

- Filter
- Wrapper
- Hybrid

Filter: The Filter methods are the one which mainly focuses on selecting subset as a preprocessing step based on some norms without considering the performance of the classifier

Wrapper: Generally, this methods are overtake whenever the classifier vicissitudes.

Hybrid: The Hybrid methods are combination of both filter and wrapper methods.

In this work, we consider the filter methods for feature selection. The following Fig.1 shows Block diagram of a network intrusion detection.

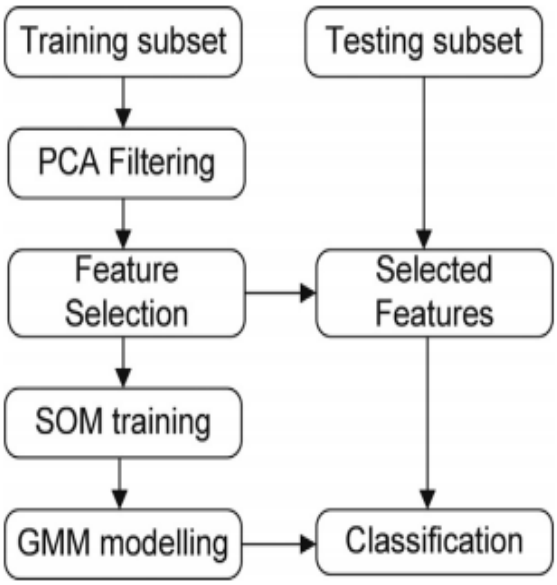

Fig-1: Block diagram of a network intrusion detection based neural network models

As we know that, Principle Component Analysis (PCA) has gained huge popularity in different applications such as image processing, data mining, pattern recognition etc. PCA is a statistical process to reduce the dimensionality of the system. The main aim of PCA is to find set uncorrelated features from the set correlated features. Often PCA is to find the internal structure of data in a way which best explains the most of the variance in the data. The proposed PCA can be working as follows.

Let X=$\{x1, \ldots\ldots\ldots xn\}$, $xi=(x_i^1, \ldots\ldots\ldots x_i^n)^T$ is the input data then we have to subtract the mean($\bar{X}$), Y=X- $\bar{X}$ where $y_i \in R^n$, $y_{i=}(y_i^1, \ldots. \; y_i^n)^T$Where i=1.....n;

PCA searches for n orthonormal vectors $u_{k=}(u_{k,\ldots\ldots}^1 u_k^n)$, **k=1.....n** Such that, $\lambda_k = \frac{1}{M}\sum_{r=1}^{n}(u_k^T y_r)^2$Vectors $u_k$and λ_k are the Eigen vectors and Eigen values correspondingly.

Although, The PCA finds the components with higher variance we cannot say higher variance attributes are the most discriminant ones. This problem can be solved by using Fisher Discriminant Ratio method.

$$FDR=\frac{(\mu_i-\mu_j)2}{\sigma_i^2+\sigma_j^2} \ldots\ldots (1)$$

Algorithm: PCA-FDR method

Step 1: Acquire some data

There are 41 features in the sample dataset will be smeared to PCA in order to select features.

Step 2: Subtract the mean as Y=X- $\bar{X}$

$$\bar{X}=\frac{\sum_{i=1}^{n} x_i}{n} \ldots\ldots (2)$$

Step 3: Find out the covariance matrix

$$Var(x) = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(x_i-\bar{X})}{(n-1)} \ldots. (3)$$

Step 4: Compute Eigen values and Eigen vectors of the covariance matrix.
Step 5: Forming features by considering principle components.
Step 6: Acquire the new data

3.2 Classification using SOM

Self-organizing maps (SOM) are also called as Kohonen's [12] Self-organizing maps (SOMs).SOM [13] is an unsupervised learning technique. The SOM is a powerful technique as it provides effective way of classifying dataset. Also SOM sanctuaries topological mappings between representations, a feature which is normal or abnormal for [14]network data. SOM is one of the category of competitive learning network. The SOM Fig-2specifies that mapping between high dimensional space to regular two dimensional space. The SOM can be used several applications like data compression, pattern recognition etc. Although some intrusion detection systems uses SOM as a classification technique in order to anomalies from normal traffic. The structure of SOM can be single feed forward network, where each source node of the input layer is connected to the output neurons The SOM algorithm can be as follows.

Let $X \epsilon R^n$ is n-dimensional data and the model vector isω_i. For every input instance v, the Best Matching Unit (BMU) can be

$$\| \omega_i - v \| \leq \| \omega_j - v \| \ldots\ldots (4)$$

The General SOM topology can be showed in Fig.2

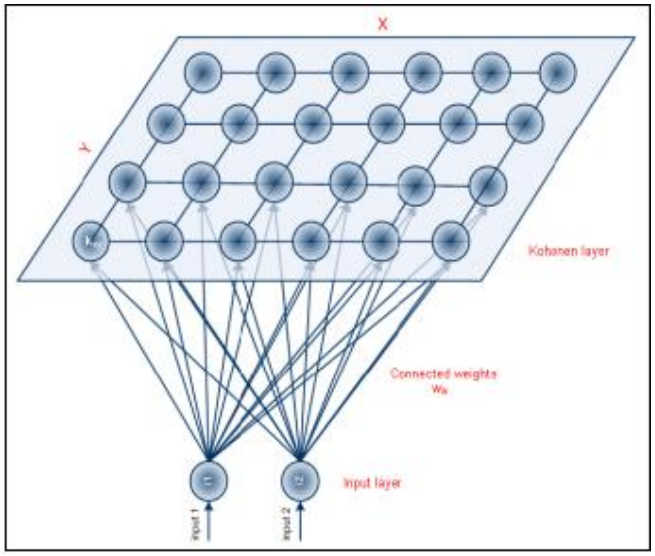

Fig-2: General SOM topology

4. Results & Conclusion

4.1 Results

The developed method can be examined by training and test datasets provided by the NSL KDD dataset. The classification performance has been examined by three parameters such as accuracy, specificity and sensitivity. The sensitivity can be calculated by using,

$$\text{Sensitivity}=\frac{TP}{TP+FN} \quad \ldots\ldots (5)$$

$$\text{Specificity}=\frac{TN}{TN+FP} \quad \ldots\ldots (6)$$

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad \ldots (7)$$

In Eqs. 5-7, TP, TN, FP, and FN are the True Positive, True Negative, False Positive, False Negative correspondingly.

Table-1: Comparison of Classification results

Method	features	Accuracy	Sensitivity	Specificity
PSOM+FDR[15]	9	0.89±0.05	0.98 ±0.06	0.77 ± 0.06
PSOM+PCA[15]	15	0.9±0.05	0.97±0.05	0.80±0.08
PSOM+FDR+PCA	**8**	**0.90**	**0.97**	**0.93**

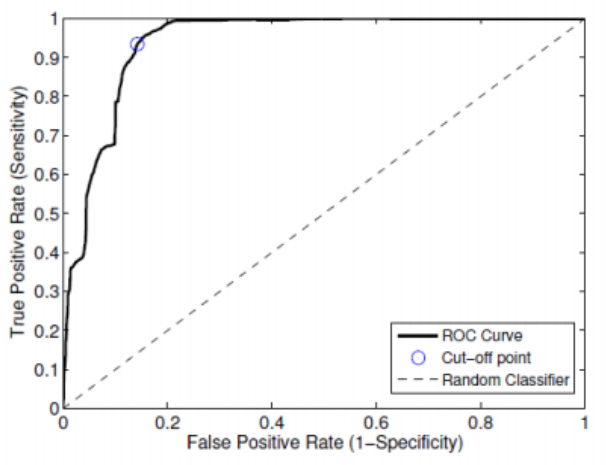

Fig-3: ROC curve for normal/attack classification using PCA as Feature selection method

5. Conclusion

In this work we designed a network intrusion system based on PCA and SOM. The Table-1 shows that our results is showing best among other methods.

6. References

[1] P. Ravi Kiran Varma and M. B. Subrahmanyam, "AdaptiveReorientationMethodforPerformanceEnhancement in Network Firewalls," *International Journal of Computer Applications,* vol. 80, no. 14, pp. 31-36, Oct 2013.

[2] A. George, "Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM," *International Journal of Computer Applications,* vol. 47, no. 21, pp. 5-8, June 2012.

[3] P. Ravi Kiran Varma and V. Valli Kumari, "Feature Optimization and Performance Improvement of aMulticlass Intrusion Detection System using PCA and ANN," *International Journal of Computer Applications,* vol. 44, no. 13, pp. 4-9, April 2012.

[4] B. Neethu, "Adaptive Intrusion Detection Using Machine Learning," *IJCSNS International Journal of Computer Science and Network Security,* vol. 13, no. 3, pp. 118-124, March 2013.

[5] C. Kruti, S. Bhavin and K. Ompriya, "Intrusion Detection System using Self Organizing Map: A Survey," *International Journal of Engineering Research and Applications,* vol. 4, no. 12, pp. 11-16, 2014.

[6] P. Ravi Kiran Varma, V. Valli Kumari and S. Srinivas Kumar, "A Novel Rough Set Attribute Reduction based on Ant Colony Optimization," *International Journal of Intelligent Systems Technologies and Applicaitons,* vol. (In Press), 2015.

[7] P. Ravi Kiran Varma and V. Valli Kumari, "A SECURITY FRAMEWORK FOR ETHERNET BASED EMBEDDED WEB SERVER," *International Journal of Embedded Systems and Applications,* vol. 2, no. 2, pp. 17-27, June 2012.

[8] "KDD 1991 datasets,The UCI LDD Archive," Irvine, CA, USA, 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[9] H. Jyoti, "Network Intrusion Detection using Semi Supervised Support Vector Machine," *IJCA,* vol. 85, no. 9, pp. 27-31, 2014.

[10] L. Yijuan, I. Cohen, Z. Xiang Sean and Q. Tian, "Feature Selection Using Principal Feature Analysis," *ACM Multimedia,* pp. 23-29, Sep 2007.

[11] A. Srivastav and R. G. Pankaj Kumar, "Evaluation of Network Intrusion DetectionSystem using PCA and NBA," *IJARCET,* vol. 2, no. 11, pp. 2873-2881, 2013.

[12] A. B. M. C. Liberios VOKOROKOS, "INTRUSION DETECTION SYSTEM USING SELF ORGANIZING MAP," *Acta Electrotechnica et Informatica ,* vol. 6, no. 1, pp. 1-5, 2006.

[13] H. Gunes Kayacik, A. Nur Zincir and I. Heywood, "On the capability of an SOM based," in *International conference on Neural Networks*.

[14] R. Sandip Sonawane, M. Shailendra Pardeshi, D. Vipul Punjabi and B. Rajnikant Wagh, "RULE LEARNIG BASED SELF ORGANIZING INTRUSION DETECTION SYSTEM," *International Journal of Research in Advent Technology ,* vol. 1, no. 3, pp. 35-42, Oct 2013.

[15] E. D. la Hoz, E. D. La Hoz, A. Ortiz, J. Ortega and B. Prieto, "PCAfiltering and probabilistic SOM for network intrusion detection," *ELSEVIER,* vol. In Proceeding, pp. 1-11, 2015.

[16] L. C. T. a. C. D. N. Tich Phuoc Tran, "Novel Intrusion Detection using Probabilistic Neural Network and Adaptive Boosting," *IJCSIS,* vol. 6, no. 1, pp. 83-91, 2009.