Addressing Challenges for Intrusion Detection System using Naive Bayes and PCA Algorithm

Saqr Mohammed Almansob
Dept of Computer Science
BAM University
Aurangabad, (M.S) India
saqrmohammed2014@gmail.com

Santosh Shivajirao Lomte
Dept of Computer Engineering
BAM University
Aurangabad, (M.S) India
drsantoshlomte@gmail.com

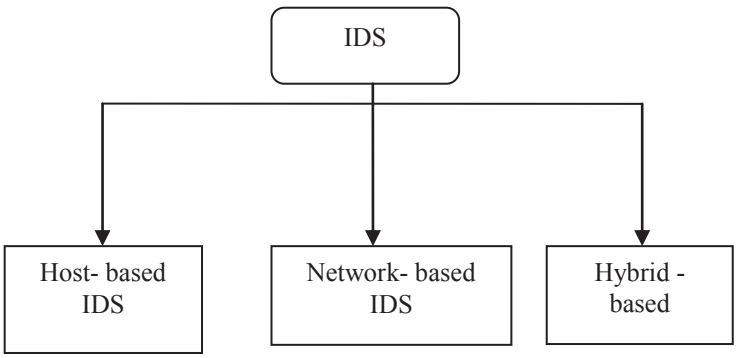
***Abstract*— Protect the network from external attacks have become the big challenge which facing networks nowadays. so, monitoring and analysis all the events over network or computer system are known as intrusion detection system. This paper proposed two approaches to addressing intrusion detection system problems. One of this approach is known as Principal Component Analysis (PCA) for feature extraction and applied Naive Bayes approach as a classification problem. so, the model applied on the KDD99 dataset. The obtained results show the increase in detection and accuracy rate as well as a decrease in false positive rate.**

Keywords*— *Intrusion Detection System, Principal Component Analysis, Naive Bayes

I. Introduction

Nowadays, the network exposed to attacks continuously which leading to stealing users files and information . For this reason, must monitor and analysis all the traffic data on the network using intrusion detection system so false positive rate one of the intrusion detection problem which generates a lot of alert over network our challenge to reduce the false positive and improve the detection and accuracy rate in network. On other hand, intrusion detection system play the role of mongering all the traffic data which grow through a network and analysis all the normal and malicious events over network. Furthermore, the attacks grew up and increased on the system which leads to the loss of a lot of important information. There are three main types of intrusion detection system. The three types of IDS description as follows

Fig 1. Types of Intrusion Detection System(IDS)



A. Host-Based intrusion detection system (HBIDS) HBIDS is a method of security control for computer systems. So, HBIDS gather all the information about network events and monitor all the traffic data in a single host. Furthermore, detect all the types of attacks on the particular host only and analysis the file system of host, loges activities, users

B. Network- based intrusion detection system (NBIDS) NBIDS used to monitor and analysis all the traffic data pass over the network and detect all types of attacks on the hosts of that network. So, the NBIDS give on alert of the attacks only whereas in HBIDS give verifying if attack was successful or not.

C. Hybride-based Intrusion Detection System (HBIDS) is a collection of HBIDS and NBIDS for monitoring all the traffic data over network as well as analysis traffic data.

The following will describe four main sections. First section about related work. Second section will explore the proposed work. Third section describe the experimental and results. Final section comprise of conclusion and reference.

II. Related work

Sandhy et al [1] have proposed Decision Tree (DT) and Support Vector Machine (SVM) for modeling intrusion detection. The objective of using Hybrid approach to increase detection accuracy and decrease computational completing .The authors applied KDD99 dataset intrusion detection for increase detection accuracy. So, The obtain results of the proposed model gave more accurate intrusion detection. Shih-Weiline et al [2] have applied Support Vector Machine (SVM), Decision Tree (DT) and Simulated Annealing (SA) for anomaly intrusion detection. The authors applied SVM and SA for choose best features to increase of anomaly intrusion detection. so, they used the KDD99 dataset in experimental, the obtained results of using SVM-SA can detect best features to increase accuracy of anomaly intrusion detection. Furthermore, the obtained results of DT-SVM get better accuracy classification and detect new attacks by decision rules from KDD99 dataset. Vladimir Bukhtoyarov et al [3] have proposed Ensemble-Distributed method to addressing all the problems of classification. So, this method appropriate for intrusion detection problem in computer network.

978-1-5090-4307-1/17/$31.00 ©2017 IEEE

Furthermore, approached conducted in KDD99 dataset, the obtained results gave higher performance for classification. Emna Bahri et al [4] have proposed Greedy-Boost method to increase faster intrusion detection. So, the authors applied this method to get high level of accuracy and increase the detection rate as well as decrease false positive rate. Furthermore, the proposed model applied on KDD99 dataset. Fangjun Kuang et al [5] have introduced KPCA and SVM with GA model for intrusion detection. The authors applied Combining Kernel Principal Component Analysis (KPCA) as preprocessor and Support Vector Machine (SVM) to reduce the high dimensionality of KDD99 dataset which applied in model for improving the classification. So, the obtained resultsof the proposed model gave higher accurate rate with faster speed. Furthermore, applied Genetic Algorithm (GA) for optimization and choose appropriate parameter for SVM classifier. The comparative results of applied between KPCA and PCA indicate that KPCA is showing better than PCA. Bhavesh Batankar et al [6] Applied Decision Tree (DT), Naive Bayes (NB) and K-nearest Napier (K-NN) to addressing the classification problems in data mining. The objective of this approaches for reduce the high dimensionality of data to improve the classification. so, the comparative results among three algorithm which applied to solve the classification problem is indicate that DT gave less error as comparative to K-NN and BN algorithms. Furthermore, Decision Tree algorithm gave higher accuracy

IV. The proposed work

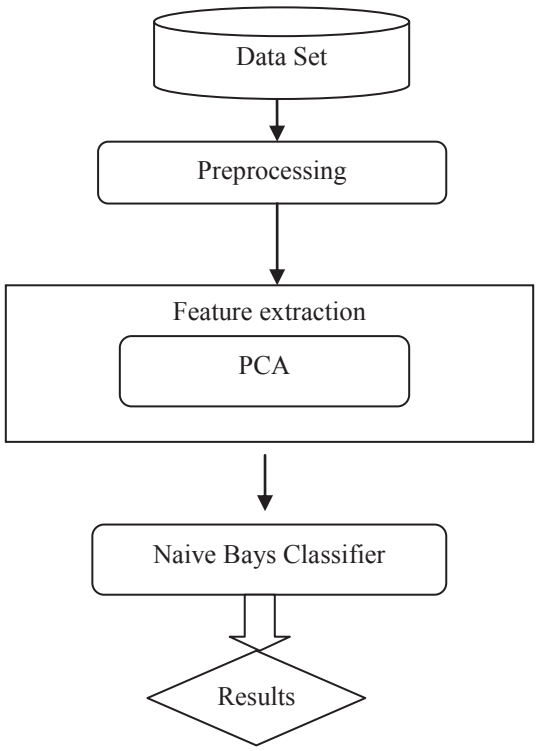


Figure 2. Generic Structure of

A. Data Set

In this experimental we used KDD99 intrusion detection data set. KDD99 has been the most wildly used data set for the evaluation of anomaly detection method. furthermore, this data has 41 features and 24 types of attacks so 14 of attacks in testing data and remaining attacks in training data. There are 4 main attacks in KDD99 are DOS, U2R, R2L and Probing so everyone of 24 types of attacks in kdd data related to one of 4 main attacks .This data have more than 4 million single connection vector each of which contains 41 features and this data is labeled as either normal or an attack.

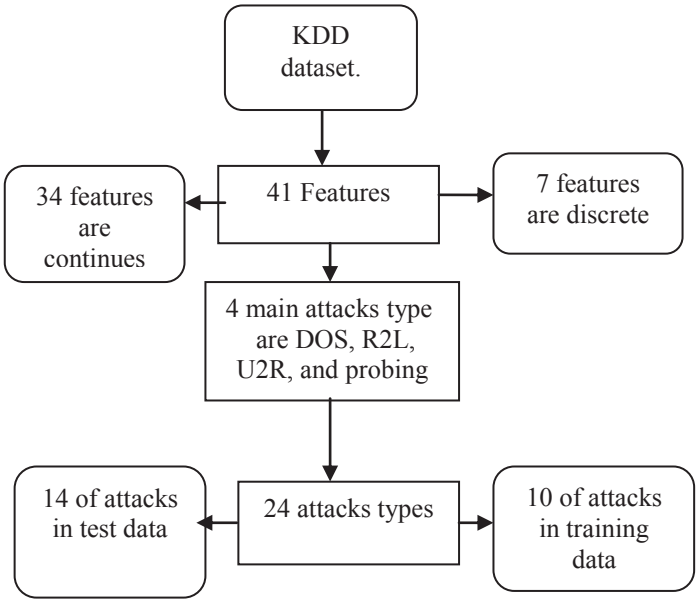


Fig 3. The classified of KDD dataset.

I. Denial of service attacks(DOS)
In this type of attack which attacks are made by sending a lot of data heavily causing slow and denial of service in which it leads the device to be out of service. There are many types of DOS attack such as back, Neptune, pod and smurf et.
II. User to Root Attacks (U2R)
These attack which the attacker start access to a normal account on the system. There are some types of U2R such as buffer_overflow, loadmodule, perl and rootkit .
III. Remote to Local Attack (R2L)
These attack happen when the attacker send packets to a machine over network but who does not have an access on that machine. There are many types of R2L such as ftp_write, guess_passwd, imap, multihop and phf.
IV. Probing Attack
These attacks happen when the attacker attempt to collect information about system and network of computers. There are a lot of probing attack types such as ipsweep, nmap, portsweep, portsweep and satan.

Table 1. The 4 types of attack in KDD dataset.

Attack name	Attack type
back	dos
buffer_overflow	u2r
ftp_write	r2l
guess_passwd	r2l
imap	r2l
ipsweep	probe
land	dos
loadmodule	u2r
multihop	r2l
Neptune	dos
nmap	probe
perl	u2r
phf	r2l
pod	dos
portsweep	probe

Some different classes used during this work. The 41 existing feature have not all same data type, some are numerical values and others are symbolic data

Table 2. The performance of the Naive Bayes

Attack type	Training time	Testing time	Accuracy
DOS	23.4	12.3	87.47
U2R	9.6	2.11	95.76
PROBE	5.2	2.9	97.12
R2L	7.3	4.13	72.16

B. Preprocessing

There are some steps for do the preprocessing the data. Firstly we import the data into matlab and plot each labeled set. secondly check for missing value and outliers data. Thirdly of preprocessing remove noise of data. Fourth Divide the data into training and testing set to build model. Furthermore, pre-process the input data into numerical data and between 0 and 1 and normalize of data finally to be numeric data whose members is number between 0an1

C. Principal Components Analysis (PCA)

Principal Components Analysis is very effective algorithm when applied to reduce high dimensionality of data in intrusion detection. Furthermore, is a attribute extraction technique that generate new attribute[7]. Principal Components Analysis can appear as next: Pci= a1x1 + a2x2 + a3x3 + a4x4 + a5x5+ a6x6..... + adxd (1)

PCi = Principal Components
Xj = Original attribute
Aj = numerical coefficient for xj

D. Naive Bayes

Naive Bayes is supervisor machine learning technique which used to classification problems. so, the KDD99 dataset contains of 41 features, 37 of features is continuous and remaining features is discrete. For this reason, Naive Bayes approach deal with continuous and discrete features[8]. Furthermore, the Naive Bayes approach is fast and can deal with big data as well as deal with missing features. Set of features X = (x1, x2, x3, x4, x5, ...,xn) set of classes Y = (y1, y2, y3, y4, y5, ...yk). Applied Naive Bayes classification to select a collection of features (x1, x2, x3, x4, x5, ...,xn) and designate those features to one of Y classes (y1, y2, y3, y4, y5, ...yk). and calculate the probabilities by utilize the training data.

$$P(C/A1, A2, A3, A4 \ldots . An) = \frac{p\left(\frac{A}{C}\right)P(C)}{P(A1,A2,A3,A4 \ldots An} \quad (2)$$

V. Experimental Analysis

The experimental implemented used MATLAB R2015a-64 bit installed on windows 7 Ultimate with the core i5 processor and 12 GB RAM . The model applied on KDD intrusion detection dataset and used 494021 instance in our experimental

A. Evaluation metrics

In this work we applied three measures for appraisal results in the experiments as follow: accuracy rate (AR), detection rate (DR), and false Positive rate (FPR). So, evaluation standard is indicate as following equations.

The evaluation standard of detection rate indicate to total numbers of attacks correctly classified which using Naive Bayes to the total numbers of attacks in the KDD99 datasets.

$$DR = \frac{\text{Number of attacks correctly classified as attacks}}{\text{Total number of attacks in kdd99 dataset}} \quad (3)$$

The evaluation standard of false positive rate indicate to total numbers of normal events which using Naive Bayes to the total numbers of attacks in the KDD99 datasets.

$$FPR = \frac{\text{Number of normal events classified as attack}}{\text{Total number of normal events in the kdd99 dataset}} \quad (4)$$

The evaluation standard of accuracy rate indicate to total numbers of classified instances which using Naive Bayes to the total numbers of instances in the KDD99 datasets.

$$ACR = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances in the kdd99 datasets}} \quad (5)$$

Number of feature	Detection Rate(DR)	False Positive Rate(FAR)	Accuracy Rate(AC)
41	79.461%	15.362%	80.143 %
35	82.131%	4.680 %	82. 312 %
30	86. 38 %	4.457 %	85. 765 %
25	88.92%	3.123 %	88. 612
20	91.67 %	3.117 %	92.126 %
15	94. 127 %	2.56 %	94.714 %
10	96. 52 %	1.16 %	96.240 %
8	98. 82 %	0.52 %	98. 53 %

VI. CONCLUSION

In this work. We have proposed two algorithms for addressing intrusion detection challenges over the network. The first algorithm applied to reduce high dimensionality and features extraction of data are known as the Principal Components Analysis algorithm and the second algorithm applied as classification problem is known as Naive Bayes algorithm. So, the model applied over KDD cup intrusion detection data set. The obtained results indicate too high detection and accuracy rate as well as lower false positive rate.

REFERENCES

[1] Sandhya P, Ajith A." Modeling intrusion detection system using hybrid intelligent systems" Journal of Network and Computer Applications. Computer Science Department, Oklahoma State University, OK 74106, USA 30 (2007) Elsevier.

[2] Shih-W, Kuo-C , Chou-Y, Zne-J." An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection" Department of Information Management, Chang Gung University, No. 259, 6 June 2012.

[3] Vladimir B and Vadim Z." Ensemble-Distributed Approach in Classification Problem Solution for Intrusion Detection Systems" Department of Information Technologies Security, Springer International Publishing Switzerland 2014.

[4] Emna B, Nouria H and Hoa N." Approach Based Ensemble Methods for Better and Faster Intrusion Detection" University of Lyon, 5, Avenue Pierre-Mendes France 69500, France Springer-Verlag Berlin Heidelberg 2011.

[5] Fangjun K , Weihong X , Siyang Z." A novel hybrid KPCA and SVM with GA model for intrusion detection"School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210018, China 2014 Elsevier .

[6] Bhavesh P, Vijay C." A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering 2014.

[7] K. Keerthi, B. Surendiran" Dimensionality reduction using Principal Component Analysis for network intrusion detection" School of Computer Science and Engineering July 2016 Elsevier.

[8] M. Ali, A. Halim , K. Gokhan" A hybrid intrusion detection system design for computer network security" 2009 Elsevier

[9] Sourabh T, Deepak G. "Data Mining Based Classification Technique for Adaptive Intrusion Detection System using Machine learning" International Journal of Advances in Engineering Sciences Vol.5, Issue 3, July, 2015

[10] Mruty U, and Manas R."NETWORK INTRUSION DETECTION USING NAÏVE BAYES" International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007.