Robust Statistical Outlier based Feature Selection Technique for Network Intrusion Detection

K. Nageswara Rao, D. Rajyalakshmi, T. Venkateswara Rao

Abstract- For the last decade, it has become essential to evaluate machine learning techniques for web based intrusion detection on the KDD Cup 99 data set. Most of the computer security breaches cannot be prevented using access and data flow control techniques. This data set has served well to identify attacks using data mining. Furthermore, selecting the relevant set of attributes for data classification is one of the most significant problems in designing a reliable classifier. Existing C4.5 decision tree technology has a problem in their learning phase to detect automatic relevant attribute selection, while some statistical classification algorithms require the feature subset to be selected in a preprocessing phase. Also, C4.5 algorithm needs strong preprocessing algorithm for numerical attributes in order to improve classifier accuracy in terms of Mean root square error. Irrelevant features in the network attack data may degrade the performance of attack detection in the decision tree classifiers. In this paper, we evaluated the influence of attribute pre-selection using Statistical techniques on real-world kddcup99 data set. Experimental result shows that accuracy of the C4.5 classifier could be improved with the robust pre-selection approach when compare to traditional feature selection techniques.

KEYWORDS: Normalization, Network security, data mining, intrusion detection, filtering.

I. INTRODUCTION

Intrusion Detection Systems (IDS) have been used to monitor network traffic thereby detect if a system is being targeted by a network attacks such as a DoS, Probe, U2R and R2L. The two main intrusion detection techniques are misuse detection and anomaly detection. Intrusion Detection concept was introduced by James Anderson in 1980**,** defined an "Intrusion attempt or threat to be potential possibility of a deliberate unauthorized attempt to access information, manipulate information, or render a system unreliable or unusable"[1]. Security of a network is always important, which monitors all network traffic passing on the segment. The following are main objectives are protecting the network against intruder's confidentiality, Integrity, Availability, Authentication and Non-repudiation. Anderson discussed a frame work investigation of intrusions and intrusion detection. In this he discussed definition of fundamental terms Risk, Threat, Attack, Vulnerability and Penetration.

Risk: Accidental or unpredictable exposure of information, or violation of operations integrity due to the malfunction of hardware or incomplete or incorrect software design.

Threat: The potential possibility of a deliberate, unauthorized attempt to:

(a) Access information
(b) Manipulate information.
(c) Render a system unreliable or unusable

Vulnerability: A known or suspected flaw in the hardware or software or operation of a system that exposes the system to penetration or its information to accidental disclosure.

Attack: A specific formulation or execution of a plan to carry out a threat.

Penetration: A successful attack; the ability to obtain unauthorized (undetected) access to files and programs or the control state of a computer system [2].

Researchers have developed two general categories of intrusion detection techniques. In misuse detection, well-known attacks or weak spots of the system are encoded into patterns, which are then used to match evidence from run-time activities to identify intrusions. In anomaly detection, normal behavior of user and system activities are first summarized into normal profiles, which are then used as yardsticks so that run-time activities that result in significant deviation are flagged as probable intrusions. Many current intrusion detection systems (IDSs) have included both misuse and anomaly detection components, and are generally complex and monolithic.

Firewalls and IDS:

An Intrusion detection system behavior is different from a firewall. Firewall monitors only at ports or services to determine whether to allow traffic or not.

Revised Manuscript Received on March 2012.

K.Nageswara Rao, Research Scholar (part time), Gitam University, Visakhapatnam. Mail-id: ksn_choudary@yahoo.com

Dr.D.Rajya Lakshmi, Professor and head, Dept of IT, Gitam University, Visakhapatnam.Mail-id: rdavvluri@yahoo.com

Prof.T.Venkateswara Rao, Professor, Dept of CSE**,** K.L University,Vijayawada.Mail-id: tv_venkat@yahoo.com

The firewall restricts the access between networks in order to prevent intrusion and does not signal an attack from inside the network. An Intrusion detecting system evaluates a suspected intrusion once it has taken place and signals an alarm. An IDS also watches for attacks that originate from within a system. The majority of attacks come from within the local network for example a buffer overflow aims to gain root access on a vulnerable web server. The owner of that server will not know that they have been compromised until it is too late. Here and IDS comes into play. An Intrusion detection system can be looked at as keeping the firewall honest. It does not take the place of the firewall. For that mater an Intrusion detection system does not block or stop traffic it only alerts a person when dangerous traffic is seen. Put another way an Intrusion detection system is a layer of defense in depth. The firewall blocks traffic destined for ports that do not have legitimate public services on them and the Intrusion detection system alerts when something potentially dangerous is seen using one of the ports that is allowed through the firewall. An Intrusion detection system can also be configured to notify when malicious traffic is seen within the LAN. For example, a network aware virus will often try to attack other computers on the local LAN as well as on the Internet. An Intrusion detection system should start on this behavior while the firewall will simply block or allow traffic depending on the port the traffic is traversing. Firewalls are used to block very broad ranges of traffic. A firewall of a web server might stop all traffic except HTTP, HTTPS, telnet and SSH. This means that even though the server system may be vulnerable to a buffer overflow type of attack in a DNS service on the server it is protected from anyone outside the firewall exploiting that vulnerability because the attacker cannot get to the vulnerable service. An Intrusion detection system allows all traffic through and only alerts when known dangerous traffic is seen[3].

Intrusion and intrusion detection

In order to detect intrusion, data collection plays an important role. On the one hand, information can be collected through operating system or applications. On the other hand, network based data collection is to detect intrusions by monitoring network traffic. Intrusion detection can also be classified into two fields: misuse detection and anomaly detection [4].

Host-based detection vs. network based detection

Intrusion detection tools can be classified into network-based or host based intrusion detection. Host-based systems analyze data from the operating system or applications subject to attack. Network-based systems look for sign of intrusions from network traffic being monitored.

Host-based detection

Modern operating systems provide auditing, logging and performance monitor to detect intrusion. Most host based systems collect data continuously as the system is operating, but periodic snapshots of the system state can also provide data that has the potential to reveal unexpected changes. Anyway, host-based detection can not select auditing to detect intrusion owing to the lack of necessary information about the operating system. Unselective logging of messages actually may incur extra auditing overhead and analysis burdens. And selective logging is hard to determine without required knowledge and computation.

Network-based detection

Network-based data collection has the advantage that a single sensor, properly placed, can monitor a number of hosts and can look for attacks that target multiple hosts. With the ease of construction, network monitoring is introduced in many commercial intrusion detection systems.

Anomaly detection vs. misuse detection

Anomaly detection is based on the assumption that misuse or intrusive behavior deviates from normal system use. Misuse detection seeks to discover intrusions by precisely defining the signatures ahead of time and watching for their occurrence. Anomaly detection, or not-use detection, differs from signature detection in the subject of the model. Instead of modeling intrusions, anomaly detectors create a model of normal “use” and look for activity that does not conform. Deviations are labeled as attacks because they do not fit the “use” model, thus the name, not-use detection. The difficulty in creating an anomaly detector is creating the model of normal “use.” The traditional method, called statistical or behavioral anomaly detection, selects key statistics about network traffic as features for a model trained to recognize normal activity.

Signature detection, also known as misuse detection, attempts to identify events that misuse a system. Signature detection is achieved by creating models of intrusions. Incoming events are compared against intrusion models to make a detection decision. When creating signatures, the model must detect an attack without any knowledge of normal traffic in the system. Attacks and only attacks should match the model; otherwise false alarms will be generated. The simplest form of misuse detection uses simple pattern matching to compare network packets against binary signatures of known attacks. A binary signature may be defined for a specific portion of the packet, such as the TCP flags.

For instance, an attack signature for the land attack would match packets that had the SYN flag set and had the same source and destination IP. The remaining content of the packet is irrelevant. The signature detection method is good at detecting known attacks. A well crafted signature will always detect the attack it represents. However, other packets may match the signature and generate false alarms. Signature systems are typically easily customizable and knowledgeable users can create their own signatures. Poorly formed signatures, however, are dangerous because they cause false alarms [5].

II. RELATED WORK

This section discusses the related works on IDS, Existing Preprocessing and Classifiers techniques applied on the IDS data. There are many research papers published regarding the preprocessing as well as classifiers in order to detect the intrusions in the network dataset. Most of them suffering with false positive type of errors while classifying the attacks. The following are the some of the related works suffering with low accuracy and false positive errors.

Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad et at[6] proposed two outlier detection techniques in statistical approach, linear regression and control chart techniques. The experimental results indicate that the control chart technique is better than that linear regression technique for outlier data detection. But this approach does not give effective results if the data contains missing values.

Dianhong Wang et al. [7] Proposed Improved attribute selection measure called ave rage gain,which penalizes the attribuâtes with many values by dividing the number of attributes values. But This méthode does not Handler missing or error values.Even method does not work when the data contains numerical values.

Faizal M. A., Mohd Zaki M., Shahrin S., Robiah Y, Siti Rahayu S., Nazrulazhar Bet al. [8] By using real time network traffic data, simulation data from DARPA99 and data from the experimental setup, the research has concluded that any connection that exceed the threshold value of 3 within 1 second is considered as an abnormal activity for fast attack detection. This system suffers with predefined threshold value which gives constant results for bounds calculation.

Yue Zhang et al [9] propose a new algorithm for outlier detection based on offset, and makes a new definition for outlier. This detection algorithm based on clustering analysis. Also this new algorithm for outlier detection based on offset has some problems to further study due to the limited conditions. For example, before modeling cluster, it needs to give the deviation threshold.

Peng Yang et al[10] propose a KNN based outlier detection algorithm which is consisted of two phases. Firstly, it partitions the dataset into several clusters and then in each cluster, it calculates the Kth nearest neighborhood for object to find outliers. This outlier detection approach depends on K initial value.

III. PROPOSED ARCHITECTURE

Proposed Research work introduces a new framework for offline analysis as shown in fig 1. For network intrusion detection. In this framework KDD99cup [11] dataset is given to Preprocessing stage which includes robust statistical techniques for both outlier detection as well as effective feature selection.

Algorithm used in this approach:

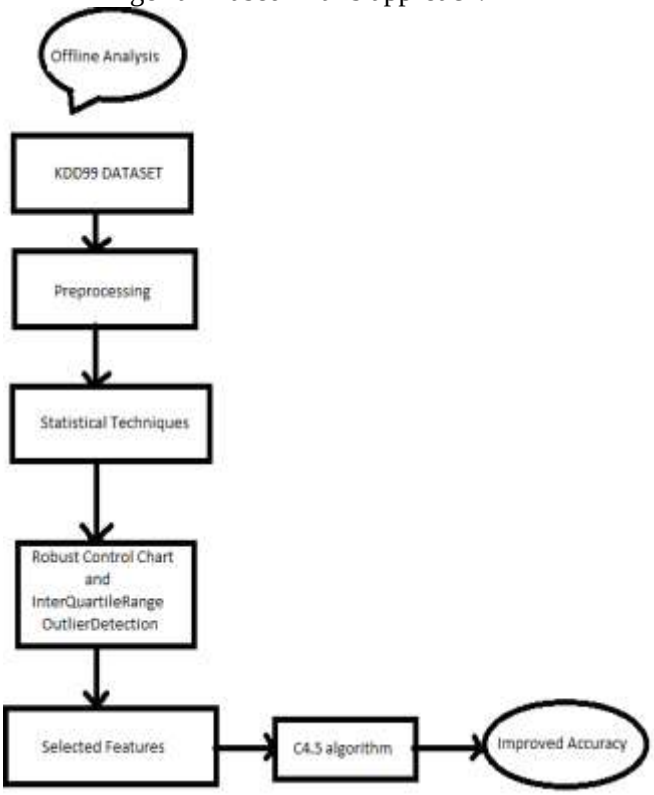


Fig 1: Proposed Architecture for Robust Feature Selection Technique.

OFFLINE ANALYSIS:

KDD CUP 99 DATA SET DESCRIPTION

Since 1999, KDD'99 [11] has been the most wildly used data set for the evaluation of anomaly detection methods. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The KDD attacks fall in one of the following four categories:

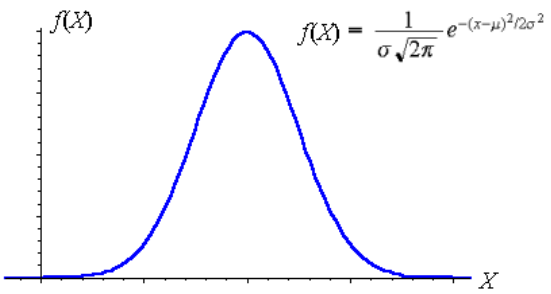
1) Denial of Service Attack (DoS): is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.
2) User to Root Attack (U2R): is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.
3) Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.
4) Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

Steps involved in Preprocessing Stage:

1) Loading dataset.
2) Apply Normality Test to each Numerical Attributes.
3) If the normality test results deviates more than the regression line, Count the high deviation points above the regression line and then attribute is passed to RIQRCC approach.
4) High deviated numerical attribute is applied to RIQRCC (Robust Inter Quartile Range Control Chart) technique.
5) If the high deviation points in the normality test are more than the outlier points in RIQRCC then that attribute is excluded in the training dataset.
6) This process is repeated to all the attributes in the dataset.

i) Normality Test:

In preprocessing stage dataset is given to statistical procedure for analyzing the data attributes. In this phase src_bytes and dest_bytes is given to Normal Probability distribution. A random variable *X* whose distribution has the shape of a normal curve is called a normal random variable**.**



Normal Curve

This random variable *X* is said to be normally distributed with mean μ and standard deviation σ if its probability distribution is given by

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

ii) RIQRCC procedure:

In this procedure first mean, standard deviation, variance is calculated to each numerical field.
For each Numerical Attribute the following steps follow.

1 Attribute values are arranged in ascending order.
2 Find the median value and represent it as M.
3 Divide the data into two partitions excluding the median.
4 Apply each partition to the following procedure
Calculate the standard deviation for each attribute as

$$\sigma = \sqrt{n\bar{p}(1-\bar{p})}$$

Where $\bar{p} = \sum_{i=1}^{k} x_i / \sum_{i=1}^{k} n_i$

x_i is the number of attacks in the attribute i that corresponds to class type anomaly.

n_i is the total number of the attack instances in the dataset that corresponds to class type anomaly.

Now, the upper action line (UAL) or control limit (UCL) may be calculated:

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

Similar method may be employed to calculate the upper warning line:

$$UWL = n\bar{p} + 2\sqrt{n\bar{p}(1-\bar{p})}$$

After calculating the UAL and UWL values the data points in each partition which satisfies UAL and UWL are given to IQR for outlier detection.
5 Find lower quartile and upper quartile ranges to the lower and upper half partitions in IQR.
6 Check the data points which doesn't satisfies IQR as outlier points.

iii) C4.5 algorithm:

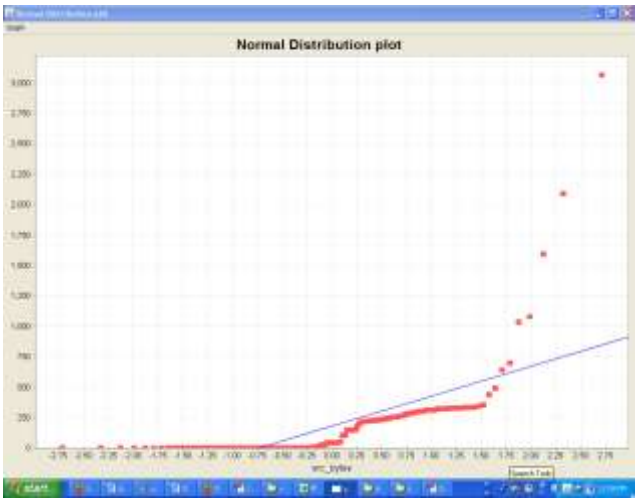
Just like Classification and Regression Tree, the C4.5 algorithms recursively visits each node, selecting the optimal split, until no further splits are possible. The steps of C4.5 algorithm for growing a decision tree is given below:

- Choose attribute for root node by using attribute selection measure Gain Ratio.
- Create branch for each value of that attribute.
- Split cases according to branches.
- Repeat process for each branch until all cases in the branch have the same class or all attributes are processed.

IV. EXPERIMENTAL RESULTS

Normality Test Results:
Normal Distribution: mean = 0.0 stdev = 1.0
Input: C5 src_bytes
Type: Probability density

X	P(X)
0.0	0.398942
232.0	0
199.0	0
0.0	0.398942
287.0	0
334.0	0
18.0	$1.75875 \cdot 10^{-71}$
8.0	$5.052271 \cdot 10^{-15}$
0.0	0.398942
303.0	0
0.0	0.398942



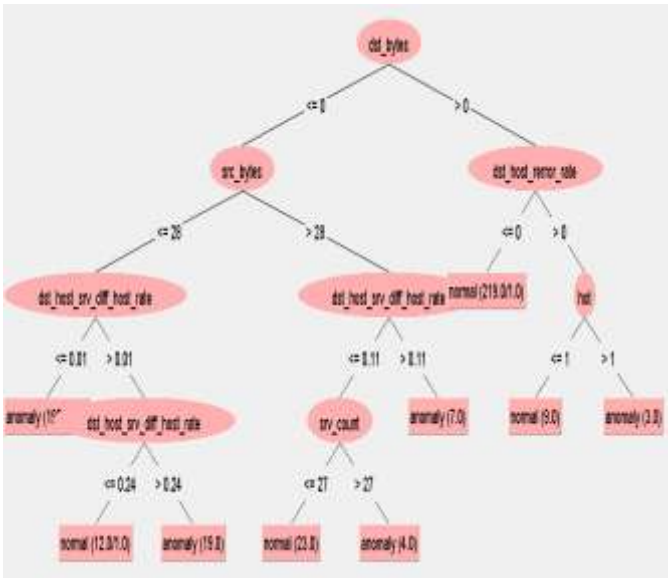
In the above normal curve the data points which are closer to the regression line are chosen to be most relevant values in the intrusion detection. In this stage we can remove some of the data values which are away from the regression line as outliers. Like this approach dest_bytes other numerical fields are applied for preprocessing.

Normality Test
Input variable: C5 src_bytes

Normality Test :
Sample Size: n = 149
Significance: 0.01
Correlation Coefficient: r = 0.67505
Critical Value: 0.98525

Critical Values are a way to save time with hypothesis testing. We don't really have to look up the probability of getting a particular value in order to verify it is less than 1% likely. Since r value less than Critical Value we can include this Src_bytes for Intrusion Analysis i.e this attribute is relevant for Intrusion Detection.

After Preprocessing, the dataset is given to C4.5 algorithm then the following results will be displayed.



Existing Info Gain Attribute with C4.5:

Correctly Classified Instances	136	91.2752 %
Incorrectly Classified Instances	13	8.7248 %

TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	CLASS(Attack Type)	C4.5 Algorithm
0.971	0.1	0.893	0.971	0.931	Anomaly	Robust Preprocessing
0.899	0.075	0.912	0.899	0.905	Anomaly	Existing Preprocessing
0.925	0.101	0.914	0.925	0.919	Normal	Existing Preprocessing
0.9	0.029	0.973	0.9	0.935	Normal	Robust Preprocessing

Proposed Attribute Selection with C4.5:

Correctly Classified Instances	139	93.2886 %
Incorrectly Classified Instances	10	6.7114 %

V. CONCLUSION AND FUTURE WORK

This Proposed work effectively identifies attacks with different types based on the relevant network features extracted using the robust feature selection algorithm. This proposed algorithm outperformed well in order to remove the outliers or noise using statistical approaches. Experimental results shows that robust statistical feature selection gives 99.36% attack detection rate when compare to other feature selection techniques.

The only limitation in this research work is implementing correct attribute selection measure in C4.5 decision tree algorithm. In future this work is extended to implement robust C4.5 algorithm to get more attack classification rate than 99.36%.

REFERENCES

[1] Anderson. J. P. "Computer Security Threat Monitoring and Surveillance." Technical Report, James P Anderson Co., Fort Washington, Pennsylvania, 1980.

[2] Shaik Akbar, Dr.K.Nageswara Rao, Dr.J.A.Chandulal "Intrusion Detection System Methodologies Based on Data Analysis" International Journal of Computer Applications (0975 – 8887) Volume 5– No.2, August 2010

[3] Daniel Owen "Network-Based Intrusion Detection Systems in the Small/Midsize Business" November of 2005, http://danielowen.com/NIDS

[4] Lixin Wang "**Artificial Neural Network for Anomaly Intrusion Detection**" http://www.cs.auckland.ac.nz/courses/compsci725s2c/archive/termpapers/725wang.pdf

[5] **Kumar Das" Protocol Anomaly Detection for Network-based Intrusion Detection** "GSEC Practical Assignment Version 1.2f (amended August, 13, 2001)

[6]A Comparative Study for Outlier Detection Techniques in Data Mining Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad(7-9 June 2006).

[7] Dianhong Wang **"An Improved Attribute Selection Measure for Decision Tree Induction"** FSKD,2007 fourth international conference.

[8] Threshold Verification Technique for Network Intrusion Detection System Faizal M. A., Mohd Zaki M., Shahrin S., Robiah Y, Siti Rahayu S., Nazrulazhar B,IJCSIS VOL2NO1(JUNE 2009).

[9] yue zhang,Jie Liu o.song,"A NEW ALGORITHM FOR OUTLIER DETECTION BASED ON OFFSET", 2009 FITth international conference on information assurance and security Chengdu **"Combining Classifier based on Decision Tree"(** 18-20 Aug. 2009)

[10] KNN Based Outlier Detection Algorithm in Large Dataset Peng Yang Chongqing University of Arts and Science Chongqing, China llylab@21cn.com

[11] kdd.ics.uci.edu/databases/**kddcup99**/**kddcup99**.html

AUTHORS PROFILE

K.Nageswara Rao is an Associate Professor and Head in the Department of Computer Science & Engineering, Mother Teresa Institute of Science and Technology, Sathupally, Khammam (Dt). Mr.Rao received M.Sc (Computer Science) from Bharatidasan University, Tiruchy, 2000, M.Tech (Computer Science & Engineering) from Bharath University, Chennai, 2005 and currently pursuing Ph.D (Computer Science & Engineering) - Part-Time from GITAM University.

Dr. D. Rajya Lakshmi is Professor & Head, in the Department of Information Technology, GIT, GITAM University, Vishakhapatnam. Dr. D. Rajya Lakshmi completed B.E (Electrical) Degree from Andhra University in 1992, M.Tech (Computer Science & Engineering) from Andhra University in 1995 and received Ph.D (Computer Science & Engineering) from JNTUH.

Dr. T.V. Rao is currently working as Professor in the department of Computer Science & Engineering in K L University, Vijayawada. Dr. Rao completed B.E (Electronics & Communication Engineering) from Andhra University, Vishakhapatnam in 1977, M.Tech (computer Science & Engineering) from PSG College of Engineering, Coimbatore in 1979 and Ph.D (computer Science & Engineering) from Wayne State University, Detroit, USA, 1992.