Privacy Preservation Approach using K-Anonymity Chinese Remainder Theorem for Intrusion Detection

[1]Sanjeevaiah Kuraganti , [2]Jeevana Jyothi. P, M.Tech
[1]*M.Tech Student, VVIT* , [2]*Asst. Prof, VVIT*

Abstract – *Privacy preservation is vital for machine learning and data mining, but measures created to protect financial information sometimes bring about a trade off: reduced utility of the workout samples. This work introduces a privacy preserving approach that could be put on decision-tree learning, without decrease in accuracy. It describes a procedure for the preservation of privacy of collected data samples if information of one's sample database continues to be partially lost. Existing approach will not work well for sample datasets with low frequency, or if low variance within the distribution of every samples. This procedure converts the first sample data sets towards a category of unreal data sets, which actually the unique samples couldn't be reconstructed with no entire team of unreal data sets. Existing approach doesn't provide privacy toward the selected attributes resulting from miss classification error, also existing attribute selection measures doesn't give optimal selection gain values. Proposed system will gives optimal attribute selection measures using improved c45 algorithm in addition to privacy on attributes. In this particular proposed implementation a fresh filtering technique for preprocessing the network attacks and an improved algorithm when it comes to the classification of KDDCUP 99 dataset. Proposed decision tree algorithm is undoubtedly an optimization method utilized for fine- tuning of a given features whereas random forest, a highly accurate classifier, is created here for various kinds of attacks classification. Proposed work will concentrate more on true positive rate compare to existing approaches. The approach works with other privacy preserving approaches, for instance cryptography, for really protection. Proposed work will concentrate more on true positive rate compare to existing approaches.*

Keywords- NID, PRIVACY PRESERVING.

1. INTRODUCTION

Modern computer networks will have to be utilized with relevant security techniques so that you could protect the information resources handled near them. Intrusion detection systems (IDSs) are integral aspects of any well configured and handled computer network systems. An IDS is basically a various software and hardware components, very effective at evaluating several activities within a network and analyze them for problems of security threats. There really are two major approaches to intrusion detection: anomaly detection and misuse detection. Misuse detection uses patterns of popular intrusions to enhance and identify unlabeled data sets. Actually, many commercial and open source intrusion detection systems are misuse based. Anomaly detection, however, is comprised of building models from normal data which might be made use to detect variations among the list of observed data seen from the normal model. Advantageous with anomaly detection algorithms is due to can detect new varieties of attacks that could deviate seen from the normal behavior. Within this project, various supervised learning algorithms, particularly decision trees as stipulated in the ID3, J48, and Naive Bayes algorithms are explored for network intrusion. Intrusion detection happens to be the art of detecting the break-ins of malicious attackers. Today, computer security has steadily grown in importance utilizing widespread utility of the world wide web. Firewalls are normaly made use to prevent attacks from occurring. Antivirus and anti-spyware programs may help individuals to remove previously existing automated attacks out of your computer. Control access limits physical and networked utility of a working laptop or computer. However, significant part in organizing secure technique is to own a technique or another to research the activity on top of the computer and figure out whether an attack continues to be launched from the computer. An incredibly structure is called an intrusion detection system. This project uses Naive Bayes, a determination Tree algorithm to discover the relative strongest and weakest points of making use of these approaches. The aim is in most cases to provide an assessment of one's performance for such algorithms which supply somebody who wishes to use a single among those approaches to learn how accurate the approach is and under what conditions it functions well. Moreover, a fresh evaluation technique will certainly be considered. Accuracy can easily be evaluated effectively with the use of Receiver Operating Characteristic (ROC) curves. Cost curves can indicate the stipulations under that's an issue algorithm works fine.

DATASET DESCRIPTION

The simulated attacks were categorized, as stated by the strategies and objectives of a given attacker. Each attack type opens into either of these two four main categories these would be attacks that are used as a a part of: The
1998 ID evaluation, The 1999 ID Evaluation training data, as well as having the 1999 ID Evaluation test data.
Attacks regarded as be New in the year 1999 is a category of people who didn't appear in 1998 as well as
1999 training data and so are denoted because of that.

1) Denials-of Service (DoS) attacks hold the goal limiting or denying services provided to the owner, computer or network. A standard tactic would be to severely overload the targeted system. (e.g. apache, smurf, Neptune, Ping of death, back, mail bomb, udpstorm, SYNflood, etc.).

2) Probing or Surveillance attacks hold the aim of gaining knowledge of this very existence or configuration regarding a computer system or network. Port Scans or sweeping regarding a given IP-address range typically fall in this category. (e.g. saint, ports weep, mscan, nmap, etc.).

3) User-to-Root (U2R) attacks provide the goal gaining root or super-user access throughout the particular computer or system on which the attacker previously had user level access. These would be attempts by way of a non-privileged user in order to increase administrative privileges (e.g. Perl, xterm, etc.).

4) Remote-to-Local(R2L) attack is undoubtedly an attack wherein an individual sends packets to your machine during the internet, which is something user lacks admittance to as a way to expose device vulnerabilities and exploit privileges which a local user is sure to have toward the computer (e.g. xclock, dictionary, guest password, phf, send mail, xsnoop, etc.).

Denial of Service Attacks

A denial of service attack can be an attack an affliction where the attacker makes some computing or memory resource too busy or too already at maximum handle legitimate requests, or denies legitimate users admission to a machine. There are several various denial of service (or DoS) attacks. Some DoS attacks (just like a mail
bomb, Neptune, or smurf attack) abuse an absolutely legitimate feature. Others (teardrop, Ping of Death) create malformed packets that confuse the TCP/IP stack of this very machine that really is aiming to reconstruct the packet. Still others (apache2, back, syslogd) benefit from bugs within a particular network daemon. The subsequent sections describe in depth each one of the Denial of Service attacks which are found in the 1999 DARPA intrusion detection analysis.

User to Root Attacks

User to Root exploits really are a importance of exploit an affliction where the attacker begins with admittance to a standard user account toward the system (perhaps gained by sniffing passwords, a dictionary attack, or
social engineering) that is ready to exploit some vulnerability to put on root admission to the buzzinar viral sales funnel. There are a number of many kinds of User to Root attacks. Some of the most common would be the buffer overflow attack. Buffer overflows occur every time a program copies an excessive amount data towards a static buffer without checking to be sure that the comprehensive data will fit into the

User to Root Attacks

Eject Description: The Eject attack exploits a buffer overflow will be the binary distributed with Solaris 2.5. In Solaris 2.5, removable media devices that lack an eject button or removable media devices that are managed by Volume Management make use of the eject program. As
a consequence of insufficient bounds checking on arguments among the volume management library, libvolmgt.so.1, you will be able to overwrite the interior stack space of a given eject program. If exploited, this vulnerability may be used in order to increase root access on attacked systems.

3.2 DATASET FILE FORMAT S

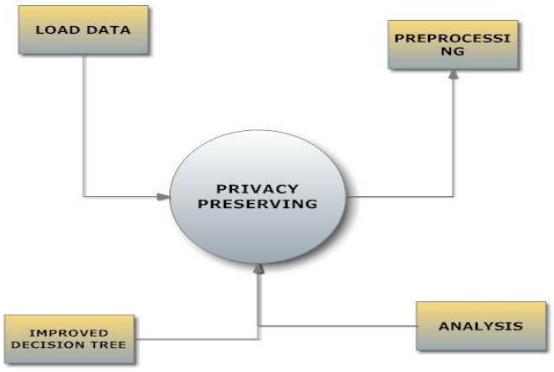
In this project two types of file formats are us ed. They are 1) CSV 2)ARFF .1) CSV: It means Comma Separated Value. This format is obtained using MS- Excel. KDD99 dataset is loaded into Excel after which it has been saved having extension of csv. 2) ARFF: It refers to Attribute Relation File Format. An file can be an ASCII text file that describes an array of instances sharing an arrangement attributes. ARFF files were developed from the Machine Learning Project for the Department of Computer Science of The college of Waikato for utilization when using the Weka machine learning software.ARFF files have two distinct sections. The very first section will be the Header information, which is certainly followed the comprehensive data inside the ARFF Header Part.

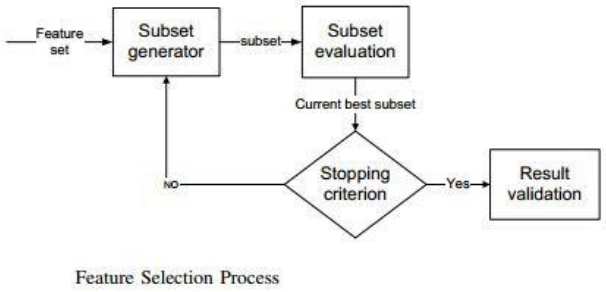
SAMPLE KDD DATASET:

@relation kdddata
@attribute 0 numeric
@attribute udp {udp,tcp,icmp}
@attribute private
{private,domain_u,http,smtp,ftp_data,ftp,eco_i,oth er,auth,ecr_i,IRC,X11,finger,time,domain,telnet}
@attribute SF {SF,RSTR,S1,REJ}
@attribute 105 numeric

@attribute 244 numeric
@attribute 0 numeric
@attribute normal.
{normal.,snmpgetattack.,named.,xlock.,smurf.}
@data
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,1,1,0,0,0,0,1,0,0,255,254,1,0.01,0,0,0,0,0,0,n
ormal.
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,1,1,0,0,0,0,1,0,0,255,254,1,0.01,0,0,0,0,0,0,n ormal.
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,2,2,0,0,0,0,1,0,0,255,254,1,0.01,0,0,0,0,0,0,s nmpgetattack.

III. PROPOSED SYSTEM





Feature Selection Process

DATALOAD: This feature is used to load the dataset for privacy preserving purpose.

PREPROCESSING:

Feature selection research has found applications in many fields where large volumes of data present challenges to effective data analysis and processing. As data evolves to be ubiquitous and abundant, new challenges arise everyday and expectations of feature selection are also elevated. Feature selection algorithms have two main components: feature search and feature subset evaluation. Feature search strategies have been widely used for searching feature space. An exhaustive search would certainly find the optimal solution; however, for a dataset of N features, a search on 2^N possible feature combinations is obviously computationally impractical. More realistic search strategies have been studied to make the problems more tractable.

PROPOSED ALGORITHMS:

Algorithm: Improved Discretization method for Decision tree.
Inp
ut:
• N, number of examples.
• Ai, continuous attributes.
• Cj, class values in training set.
. Global Threshold value
Output: Interval borders in Ai'
Procedur
e:
1. for each continuous attribute Ai in training dataset do
2. Do normalize the attribute within 0-1 range
3. Sorting the values of continuous attribute Ai in ascending order.
4. for each class Cj in training dataset do
5. Find the minimum (Minvalue) using StdDev attribute value of Ai for Cj•
6. Find the maximum (Max) attribute value of Ai for Cj.
7. endfor

8. Find the cut points in the continuous attributes values based on the Min and Max values of each class Cj.

Best Cutpoint range measure:

9. Find the conditional probability P(Cj/A) on each cut point and select the cut point with maximum probability value.

Stopping criteria:

10. If the cut point using the maximum probability value is exist and satisfies the global threshold value then it can be taken as an interval border else consider the next cut point, where information gain value and global threshold value satisfy the same point.

12. endfor

DECISION TREE ALGORITHM

The C4.5 Algorithm is the extension of ID3 algorithm .It used a mechanism of learning from large datasets .The attribute selection of the algorithm is based on an assumption the complexity of decision tree and the amount of information is represented by given attribute are closely related C4.5 expands the classify range to digital attributes. That metric standard of two-class entropy ,the most of the algorithm is based on the information entropy which is contained by produced nodal points of decision tree is least [9].The so called entropy is representative of degree of disorder of objects in the system. It is easy to understand that the smaller entropy the smaller disorder .In the other word the more sequential in the record collection, the more consistent .This is the target we seek; too .Suppose the set S is a training sample, the formula of entropy as follows:

Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value[10].design the degree of balance coefficient of a certain attribute as

Algorithm: Generate decisiontree.

Description: Generate a decisiontree from the given training-data.

Input: The training samples representedby discrete valued attribute; the set of ttributes, attributelist.

Final output: A decision tree with filtered rules. Method:

(1) createa node N;
(2) if samples are all of the same class Cls then
(3) return N as a leaf node labeled with the class Cls;
(4) if attributelist is empty then
(5) return N as a leaf node taggedwith the most common class in samples;
(6) select testattribute, the attribute among attributelist with the highest information gain;
(7) label node N with testattribute;
(8) for each individual known value ai of testattribute; (9) grow a branch from node N for the condition testattribute = ai;
(10) let si be the set of samples in samples for which testattribute = ai;
(11) if si is empty thus
(12) attach a leaf labeled with possibly the most common class in samples;
(13) else attach the node returned by Generatedecisiontree(si, attributelisttestattribute);

RESULTS:

Time taken to build model: 0.34 seconds
Time taken to test model on training data: 0.37 seconds

=== Error on training data ===

Correctly Classified Instances 5101 96.409 %
Incorrectly Classified Instances 190 3.591 %
Coverage of cases (0.95 level) 96.409 % Mean rel. region size (0.95 level) 50 % Total Number of Instances 5291

=== Confusion Matrix ===

```
   a    b   <-- classified as
2646  149 |  a = normal
  41 2455 |  b = anomaly
```

Result 2:

Enter 1: Existing Approach
Enter 2: Proposed Improved Method

2

2

IMPROVED DECISION RULES
---

src_bytes > 28 && count <= 312 && hot <= 1 &&
dst_host_same_src_port_rate <= 0.99 &&
dst_bytes > 4 &&
rerror_rate <= 0.04 ==> normal

count > 24 &&
dst_host_diff_srv_rate > 0 &&
src_bytes <= 32 ==> anomaly

dst_host_srv_rerror_rate > 0.98 &&
dst_host_srv_count <= 3 &&
dst_host_rerror_rate > 0.05 ==> anomaly

dst_host_srv_count > 155 ==> normal flag = S0 &&
dst_host_same_src_port_rate <= 0.04 ==>
anomaly

flag = REJ ==> normal flag = RSTR ==> normal flag = SH ==> anomaly

serror_rate <= 0.02 && dst_bytes <= 2638 && dst_bytes <= 2198 &&
dst_host_srv_diff_host_rate <= 0.05 &&
service = other &&
src_bytes > 92 ==> normal

serror_rate <= 0.02 &&
num_root <= 0 &&
dst_bytes > 2638 ==> normal

num_failed_logins <= 0 &&
dst_host_srv_serror_rate <= 0 &&
flag = SF &&
src_bytes > 20 &&

dst_host_serror_rate <= 0 ==> normal

protocol_type = tcp &&
dst_host_rerror_rate > 0.06 ==> anomaly
dst_host_srv_rerror_rate <= 0.02 ==> anomaly

: normal

Number of Rules : 32

Time taken to build model: 1.39 seconds
Time taken to test model on training data: 0.14 seconds

=== Error on training data ===

Correctly Classified Instances 5278
99.7543 %
Incorrectly Classified Instances 13
0.2457 %
Coverage of cases (0.95 level) 99.811 % Mean rel. region size (0.95 level) 50.1985 %
Total Number of Instances 5291

=== Confusion Matrix ===

a b <-- classified as
2793 2 | a = normal
11 2485 | b = anomaly

BUILD SUCCESSFUL (total time: 54 seconds)

2

IMPROVED DECISION RULES
---

rE39zH34JD3l6BU/1ge2+A== > 28 && i9uIdKWV8UfvkigXKmsmJQ== <= 312 && OdAEqoZx1
+uBPtez53BMYQ== <= 1 &&
EachIJKiAlhIHCBI6yyqBqZsqm13BU7JHfq7sV6+ VAM= <= 0.99 &&
7ZWOPrggp0aNBwvF8erSng== > 4 &&
4U4ZJXzu+TJcSixREJ0aiQ== <= 0.04 ==>
normal

i9uIdKWV8UfvkigXKmsmJQ== > 24 &&
Hn7bJkb3HLojXkHQ5wHGBqXYLuQYfW8+o8+1
yZ6e9kk= > 0 &&

rE39zH34JD3l6BU/1ge2+A== <= 32 ==>
anomaly

VB6scCxelvzKzztZHXfXAw== = private &&

rE39zH34JD3l6BU/1ge2+A== <= 102 ==> anomaly

alPpDIY22kSDFTrKrFD8hA== > 23 &&
7ZWOPrggp0aNBwvF8erSng== <= 1 ==> anomaly

/wEo069xQaC24NHHnM5ha7fQ8RYnAc+/M86oF
520K0s= > 0.95 &&
+uLwG2embmBNxVE5c8gTBg== = S0 ==> anomaly

l0UPpYJOFzpSbWqatLJpRh5LUoz2POLmDkBi0L
60N98= > 0.24 &&
QlWgjFonYG0j1IGIUstG0g== = icmp &&
rE39zH34JD3l6BU/1ge2+A== <= 20 ==> anomaly

SCgGE0ASnhxsVBYAjXhl2g== > 0 &&
rE39zH34JD3l6BU/1ge2+A== > 1130 ==> anomaly

Hn7bJkb3HLojXkHQ5wHGBqXYLuQYfW8+o8+1 yZ6e9kk= <= 0.63 && OdAEqoZx1+uBPtez53BMYQ== <= 19 &&
EachIJKiAlhIHCBI6yyqBqZsqm13BU7JHfq7sV6+ VAM= <= 0.78 &&
QNw3wuom2B8fRs8RXjwnJE27CzPZK9T5e22pE
qqYqc4= <= 0.06 &&
97oO35mihk8IILM+kF51bI74tcLZBbEBk8P07Pq4
tsY= > 4 ==> normal

kWxUdiW7a5x09XFUggl0LLfQ8RYnAc+/M86oF5
20K0s= > 0.98 &&
97oO35mihk8IILM+kF51bI74tcLZBbEBk8P07Pq4
tsY= <= 3 &&
pVrfb1jADDV961nUvhAOZE27CzPZK9T5e22pEq qYqc4= > 0.05 ==> anomaly

V. CONCLUSION AND FUTURE SCOPE

Within this paper, we have proposed an efficient scalable improved privacy protecting based decision tree construction algorithm which leads to high processing speed and small scale. Due to this reason, it is most appropriate for large datasets. Our suggested algorithm has numerous advantages, however the most important thing is the idea that it takes only one skip the work outs dataset for the complete construction of decision tree. This means that it significantly lowers the IO cost. Moreover, our algorithm presents a general framework which can be used for practically any existing decision tree construction algorithms and requires one unit time sorting for the numeral attribute. Hence, it reduces the sorting amount of numeral attributes and execution time of split phase in the decision tree construction process. Seen from the experimental evaluation, we've purchased a promising result, since our suggested algorithm outperforms the most present decision tree based privacy preserving techniques in terms of accuracy and execution time.

REFERENCES

[1] Wenliang Du, Zhijun Zhan. Using randomized response techniques for privacy-preserving data mining[C].The 9th ACM SIGKDD Int'l Conf. Knowledge Discovery in Databases and Data Mining, Washington, D.C., 2003.

[2] LUO Yong-long, HUANG Liu-sheng, JING Wei-wei, YAOYi-fei, CHEN Guo-liang. An algorithm for privacy-preserving Boolean association rule mining. Chinese Journal of Electronics, 2005,33(5):900-903.

[3] Zhang P, Tong YH, Tang SW, Yang DQ, Ma XL. An effective method for privacy preserving association rule mining. Journal of Software, 2006, 17(8):1764-1774.

[4] Fang Weiwei, Hu Jian, Yang Bingru. Studies on Privacy Preserving of Distributed Decision Tree Mining [A]. Computer Science, 2009 36(4):239-241

[5] Zhong Alin, Xu Fangheng. Studies on New Technology of Database Encryption [A], Journal of Henan Normal University (Natural Science), 2007 35(4) :51-53

[6] W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In KDD, pages 505– 510, 2003.

[7] A. V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 217–228, 2002.

[8] M. Kantarcioglu and C. Clifton. Privately computing a distributed k- nn classifier. In J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, PKDD, volume 3202 of Lecture Notes in Computer Science, pages 279–290. Springer, 2004.

[9] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In ICDM, pages 99–106. IEEE Computer Society, 2003.

[10] Y. Lindell and B. Pinkas. Privacy preserving data mining. In M. Bellare, editor, CRYPTO, volume 1880 of Lecture Notes in Computer Science, pages 36–54. Springer, 2000

[11] Privacy Preserving Decision Tree Learning Using Unrealized Data Sets Pui K. Fong and Jens H. Weber-Jahnke, Senior member, IEEE Computer Society. IEEE Transactions on knowledge and data Engineering, vol. 24, No. 2, February 2012, Pages 353-364