IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 24-29
www.iosrjournals.org

A Statistical Pattern Classification and Data Mining Approach through Cloud Computing and Security

V.R. Nikam Dr. G. S. Katkar, Dr.MK Umathe
Dept .Comp. Sci.& Mgt. Taywade college, Koradi Nagpur Commerce college, Nagpur
HOD Computer Science Taywade college, Koradi Nagpur Commerce college, Nagpur
Arts, sci. & R. Mokhare Taywade college, Koradi Nagpur Commerce college, Nagpur

Abstract- *Pattern recognition as a field of study developed significantly in the 1960s. It was very much an interdisciplinary subject, covering developments in the areas of statistics, engineering, artificial intelligence, computer science, psychology and physiology, among others Increasing adoption of cloud computing technology worldwide has also increased stress on its privacy and security issue. Out of many methods for prevention of penetration on cloud, statistical machine learning and data mining is gaining attention due to its accuracy. Statistical classification can be done using Naïve Bayes and J48 (ID3) algorithm on two data traces sample of ZEUS/Zbot a trogan horse using WEKA an open source machine learning and data mining tool . There are 11 attribute consider for classification on this algorithm. These classified result are cross verified from both the algorithm. The best rules for association are also identified for data mining using Apriori algorithm. Selection of attribute for proper data mining is identified using information gain as attribute evaluator and searching is done using ranking algorithm, giving best possible attribute for data mining of a particular type of attribute to be searched. Some practical implementation possibilities are also suggested in an abstract level algorithm.*

Key words: *Network traffic classification, WEKA, Pattern classification, Data mining,*

I. INTRODUCTION

Cloud computing is widely adopted recent technology. The growth of cloud computing technology has also initiated its security issue. Privacy is the most important issue for cloud computing developer. All these security issue needs to be tackled with faster and efficient way but also require having a better availability and recovery time of authorized user. The situation is not just limited to different type of cloud computing services but also include cloud computing as security services which itself is prone to security penetration.

Once the security of cloud is broken it give enormous access to an unauthorized user, some time even highly confidential information can be leaked out. So cloud computing can become a boon for us if we have a very powerful and efficient cloud security available with us. Basically these security loop hole are exploited by hacker using malware content like viruses, trogan hourse etc.

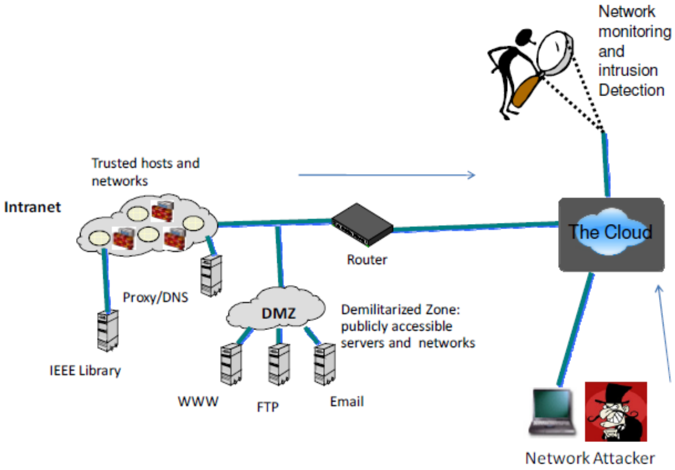

Network monitoring and intrusion Detection
Trusted hosts and networks
Intranet
Router
The Cloud
Proxy/DNS
DMZ
Demilitarized Zone: publicly accessible servers and networks
IEEE Library
WWW
FTP
Email
Network Attacker

Figure-1 : The structure of Cloud computing with intrusion detection

There are many security techniques (anomaly detection) available today like port based, packets header based, packets payload based and statistical based. Out of all these techniques statistical based or machine learning approach is replacing the other technique for a simple reason that statistical approach is a conclusion based approach and can be easily implemented in any network. Its computational time for training is more whereas implementation based on conclusion of training is very less hence can be implemented on real time application like SNORT an powerful and open source intrusion detection system.

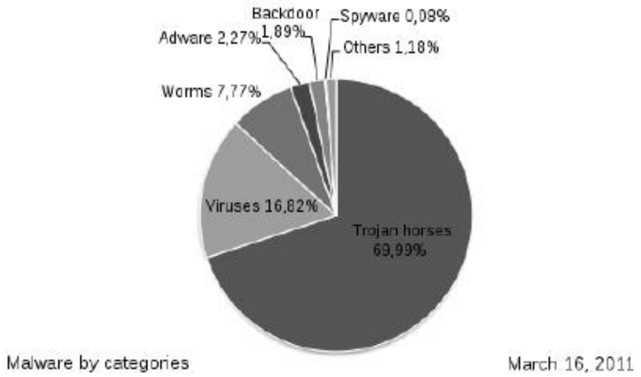
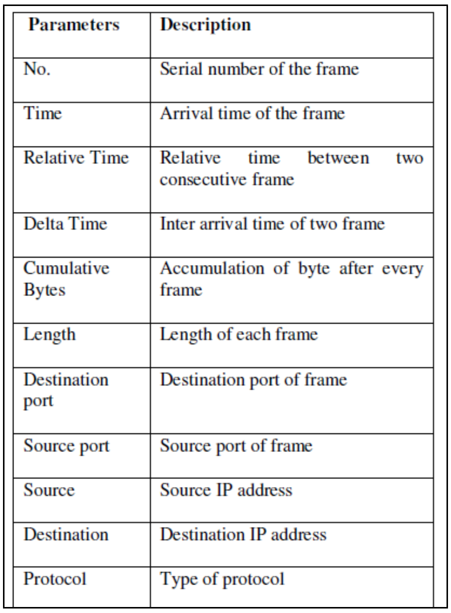


Figure-2: Distribution of malware content in internet traffic. (Courtesy Wikipedia.org)

As per the recent data 70 percent of malicious content in the internet is trogan horses. It is basically malicious developed for highly destructive application like exploitation and unauthorized gains. Trogan horses are difficult to identify once entered into any computer because unlike viruses or worm they does not replicate them self hence whatever they perform is as simple as other application. It is mainly used by hacker to gain unauthorized information from target computer. One of such trogan horse in the recent past is ZEUS/Zbot which steals the information from keystroke like password or other banking information target computer user. Zeus botnet has once millions of computer around the world with target of stilling net banking related information. It is first appearing and till 2010 it has stolen about 70million US dollar by unauthorized access to the bank account by stealing the password. Its source code is revealed in 2010 and its packet capture file or network traffic of ZEUS available in www.openpacket.org . By using tools like WEKA 3.7 (an opens so learning and data mining tool) traces of ZEUS trogan horse can be easily classified with different algorithm and parameters. The two most common and highly effective pattern classification algorithm are naïve bayes and ID3. Using these classifier in WEKA sample of ZEUS trogan horse are classified based on probabilistic and tree based model for implementation in software code in order to detect such malicious activity in the cloud on real time and take necessary measure to avoid its impact to the user. II. Statistical approach for ZEUS traffic classification Considering two different sample capture of ZEUS in the form of packet captured file are converted to its ARFF equivalent in the WEKA ARFF generator. The patterns taken into consideration are given below :

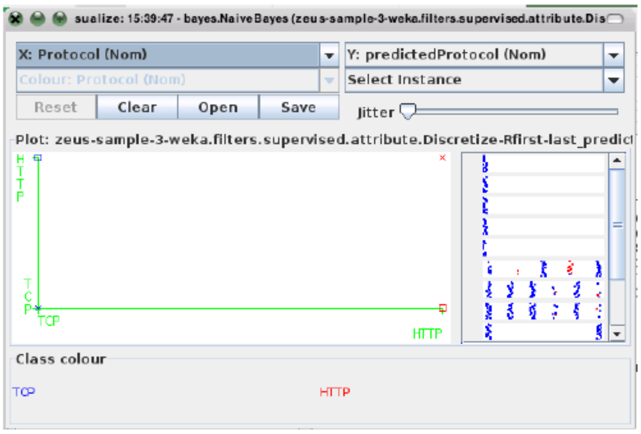


Parameters	Description
No.	Serial number of the frame
Time	Arrival time of the frame
Relative Time	Relative time between two consecutive frame
Delta Time	Inter arrival time of two frame
Cumulative Bytes	Accumulation of byte after every frame
Length	Length of each frame
Destination port	Destination port of frame
Source port	Source port of frame
Source	Source IP address
Destination	Destination IP address
Protocol	Type of protocol

The approaches of classification for this pattern are on classifier namely Naïve Bayes and ID3 (J48 in WEKA). The Naïve Bayes classifier is Probabilistic classifier based on Bayes theorem whereas ID3 algorithm is based on decision tree algorithm by considering factor like information entrop and information gain. Navie Bayes classifier is best used for supervised learning and predicts the new input according to probabilistic model. Similarly ID3, a statistical classification algorithm is used for making decision tree model based on information provided. These samples are taken into consideration and discretize first then passed on for either ID3 (J48) or Naïve B classification to generate statistical pattern classification results. It is completely supervised learning mechanism to generate training sets for anonymous input in the form of network traffic.

A. Naïve Bayes Classification

are taken into consideration with reference as protocol (HTTP, TCP) for generation of probabilistic model In this type of classifier two sample of captured ZEUS trogan horse.



sualize: 15:39:47 - bayes.NaiveBayes (zeus-sample-3-weka.filters.supervised.attribute.Dis

X: Protocol (Nom)

Y: predictedProtocol (Nom)

Colour: Protocol (Nom)

Select Instance

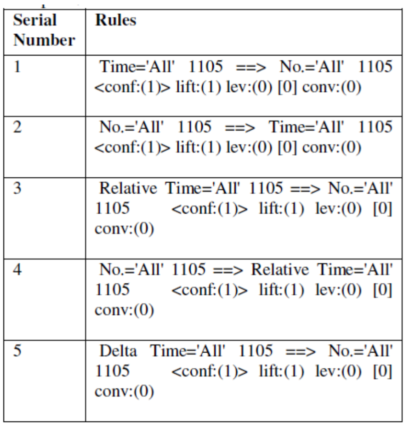
Plot: zeus-sample-3-weka.filters.supervised.attribute.Discretize-Rfirst-last_predict

Figure-3: Classification error in Naïve Bayes classifier for the given sample data set.

The outputs after the classification of two sets of ZEUS sample traffic are tabulated below.

Parameters	For Sample 1 (110 instances, 11 attributes)	For sample 2 (1105 instances, 11 attributes)
Correctly Classified Instances	110 (100%)	1100 (99.5475 %)
Incorrectly Classified Instances	0 (0%)	5 (0.4525 %)
Kappa statistic	1	0.8253
Mean absolute error	0.0222	0.009
Coverage of cases (0.95 level)	100 %	99.7285 %
classified as	a b <-- classified 100 0 | a = TCP 0 10 | b = HTTP	a b <-- classified 1088 3 | a = TCP 2 12 | b = HTTP

Bayes classification in WEKA. From the result table it can be clearly inferred that Naïve Bayes classifier is able to identify the entire sample for identification of protocol. Now using WEKA it can also associate different data with other parameter in the pattern table. The results of association of this parameter are listed below in table 3. These are stated in WEKA as best rule for learning association for data mining on the data set of sample 2.



Serial Number	Rules
1	Time='All' 1105 ==> No.='All' 1105 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2	No.='All' 1105 ==> Time='All' 1105 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3	Relative Time='All' 1105 ==> No.='All' 1105 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4	No.='All' 1105 ==> Relative Time='All' 1105 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5	Delta Time='All' 1105 ==> No.='All' 1105 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

Table-3: Association in WEKA and best data mining rule sets.

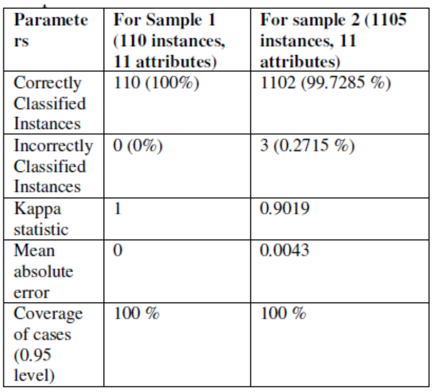
The rule used for learning association is A priori rule The interpretations of this rule are justified as for the first case time (Time) can be easily identified from serial number (No.) with conf as probability of confidence, lev for probability of leverage conv for probability of conviction and lift for probability of equability. This interpretation can be applicable for all other rules in the table 3. For data mining and select attribute play an important role. WEKA has provided feature for this. There are different attribute evaluator and search method for evaluating the attribute in the dataset and search method for identifying right attribute for out of the other parameter for data set of sample 2. The result of attribute selector is listed below.

```
Search Method:
Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 11 Protocol): Information Gain Ranking Filter
Ranked attributes:
0.088515        6 Length
0.012693        7 Destination port
0.011075        8 Source port
0.000507        10 Destination
0.000507        9 Source
0       2 Time
0       1 No.
0       5 Cumulative Bytes
0       3 Relative ime
0       4 Delta Time
```

The attribute evaluator used is Information Gain and search method used is ranking. It is clear that identifying protocol in the dataset best attribute is length then destination port and so on. The other interpretation of this attribute selector is data mining ranking for identifying protocol in the given data set. The best practice for data mining for attribute will be when number of feature is few and result for selected attribute is accurate. All this above mentioned method is need to be applied for other parameter or attribute before coming out with a software algorithm to be programmed in the network devices.

B. ID3 algorithm (J48 in WEKA)

Similar to Navie Bayes classifier two sample of captured ZEUS trogan horse are taken into consideration with reference as protocol (HTTP, TCP) for generation of probabilistic model using J48 or ID3 classifier. It is a decision tree based classifier. The outputs after the classification of two sets of ZEUS sample traffic are listed in table-4 below.



Parameters	For Sample 1 (110 instances, 11 attributes)	For sample 2 (1105 instances, 11 attributes)
Correctly Classified Instances	110 (100%)	1102 (99.7285 %)
Incorrectly Classified Instances	0 (0%)	3 (0.2715 %)
Kappa statistic	1	0.9019
Mean absolute error	0	0.0043
Coverage of cases (0.95 level)	100 %	100 %

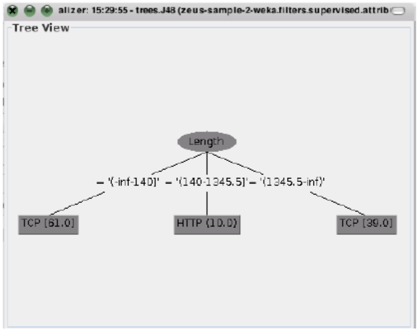


Figure-4. J48 Classified decision tree for ZEUS data trace.

The result for other properties data like association and selection of attribute for data mining is same for the data set of sample 2 provided the data set is discredited first.

II. Result

From the above performed pattern classification and data mining for identifying protocol and mine protocol in the provided data set from least feature like length and destination port. Similarly other attribute in the data set are classified and a proper mining approach is established for faster cross verification of ZEUS trogan horse and other similar internet anomalies from their data traces. After all these effort a final identification software algorithm can be implemented in the network or cloud or even cloud computing as a security services.

III. Implementation

The implementation on any network can be very easy and with very less data base support hence performance is faster and can be even on real time. Implementation using tree based approach for detecting ZEUS/Zbot is shown in the figure5 below.

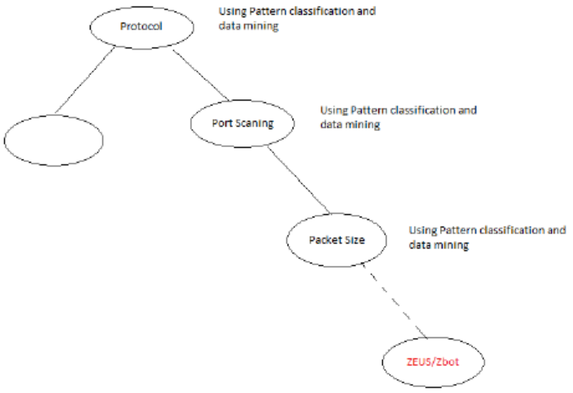

Protocol
Using Pattern classification and data mining
Port Scaning
Using Pattern classification and data mining
Packet Size
Using Pattern classification and data mining
ZEUS/Zbot

Figure-5: Detection of ZEUS trogan horse in the network before it enters a host computer.

This approach can be implemented in the intrusion detection system like SNORT or any other similar application.

IV. CONCLUSION

Using statistical pattern classification and data mining approach for intrusion detection on cloud can be implemented on the software code may be on real time. Instead of having a huge database or requirement of high computation power for pattern classification on real time network, the approach of classifying the intrusion traffic and then implementing it in the network as conclusion makes network intrusion monitoring faster and may be on real time. Intrusion detection using pattern classification of network traffic surely a promising solution for malware prone host computers mainly computer using windows based operating system.

REFERENCES

[1]. Abramson, I.S. (1982) On bandwidth variation in kernel estimates – a square root law. Annals of Statistics, 10:1217–1223.

[2]. Abu-Mostafa, Y.S., Atiya, A.F., Magdon-Ismail, M. and White, H., eds (2001) Special issue o 'Neural Networks in Financial Engineering'. IEEE Transactions on Neural Networks, 12(4)

[3]. Sebastian Zander, Thuy Nguyen, Grenville Armitage "Automated Traffic Classification and Application Identification using Machine Learning" Proceedings of the IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05).

[4]. Andrew W. Moore, Denis Zuev "Internet Traffic Classification Using Bayesian Analysis Techniques" SIGMETRICS'05, June 6.10, 2005, Banff, Alberta, Canada.

[5]. Tom Auld, Andrew W. Moore and Stephen F. Gull "Bayesian Neural Networks for Internet Traffic Classification" IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 18, NO. 1, JANUARY 2007 pp 223-239.

[6]. YU-XIN DING, MIN XIAO, AI-WU LIU "RESEARCH AND IMPLEMENTATION ON SNORT-BASED HYBRID INTRUSION DETECTION SYSTEM" Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009.

[7]. Zhuowei Li1i, Amitabha Dad and Jianying Zhou "Theoretical Basis for Intrusion Detection" Proceedings of the 2005 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY.

[8]. Duanyang Zhao, Qingxiang Xu, Zhilin Feng" Analysis and Design for Intrusion Detection System" Based on Data Mining" 2010 Second International Workshop on Education Technology and Computer Science.

[9]. Thuy T.T. Nguyen and Grenville Armitage" A Survey of Techniques for Internet Traffic Classification using Machine Learning" IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 10, NO. 4, FOURTH QUARTER 2008 pp56- 76

[10]. José V. Pagán "Improving the Classification of Terrorist Attacks" 2010 2nd International Conference on Software Technology and Engineering(ICSTE)

[11]. Snort - The de facto standard for intrusion detection/prevention, http://www.snort.org, as of August 14, 2007.

[12]. T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimize the use of Machine Learning classifiers in real-world IP networks," in Proc. IEEE 31st Conference on Local Computer Networks, Tampa,Florida, USA, November 2006

[13]. Wolpert, D.H. (1992) Stacked generalization. Neural Networks, 5(2):241–260.

[14]. Wong, S.K.M. and Poon, F.C.S. (1989) Comments on 'Approximating discrete probability distributions with dependence trees'. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(3):333–335.

[15]. Woods, K., Kegelmeyer, W.P. and Bowyer, K. (1997) Combination of multiple classifiers using local accuracy estimates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4):405–410.