Malicious Code Detection Using Naïve Bayes Classifier.

Aniket.P.Sagane[1],Prof.S.S.Dhande[2]

[1]Information Technology, SGBAU University, Amravati, Maharashtra, India
[2]Computer Science, SGBAU University, Amravati, Maharashtra, India

ABSTRACT

Today the growth of network-based services and sensitive information on networks, network security is getting more and more important. Intrusion attack has consists of serious security risk in a network environment. The growing of new intrusion types posses a serious problem for their detection. The human labeling of the available network audit data instances is usually tedious, time consuming and expensive. In this study paper, we are apply one of the efficient data mining algorithms called naïve bayes for for anomaly based network intrusion detection. The proposed technique performs better in terms of false positive rate, less cost and less Computation time.

Keywords: Network Security, Intrusion Detection, Naïve Bayes classifier.

1. INTRODUCTION

The growth of more network-based services and sensitive information on networks, network security is becoming more and more importance than ever before. Intrusion detection techniques are a only solution against computer attacks behind secure network architecture design, firewalls, and personal Screening. Despite the plethora of intrusion prevention techniques available, attacks against computer systems are still successful. Thus, intrusion detection systems (IDSs) play a vital role in network security. There many attacks are seen on confidential information such as credit card numbers, passwords, and other financial information are on the rise, going from 8 million attacks in June 2004 to over 30 millions in less than a year. One solution to this is the use of network intrusion detection systems (NIDS) that detect attacks by observing various network activities. It is therefore crucial that such systems are accurate in Identifying attacks, quick to train and generate as few false positives as possible.

1.1 INTRUSION DETECTION SYSTEM:-

An Intrusion Detection System (IDS) control on the suspicious activities in a system or patterns that may indicate system attack or misuse behavior. There are two main categories of intrusion detection techniques; Anomaly detection and Misuse detection. The former analyses the information gathered and compares it to a defined baseline of what is seen as “normal” service behavior, so it has the ability to learn how to detect network attacks that are currently unknown. Misuse Detection is based on signatures dataset for known attacks, so it is only as good as the database of attack signatures that it uses for comparison. Misuse detection has low false positive rate, but cannot detect novel attacks. However, anomaly detection can detect unknown attacks, but has high false positive rate. The specific attack are discussed in more detail in the following section

1.2 NETWORKING ATTACK:

The Following attacks were classified, according to the actions and goals of the attacker. Each attack type falls into one of the following four main categories

a)Denials-of Service (DOS) attacks are denying services provided to the user, computer or network. A common tactic is to severely overload the targeted system. (e.g. apache, smurf, Neptune, Ping of death, back, mail bomb, udpstorm, SYNflood, etc.).

b) Probing or Surveillance attacks :- The aim of gaining knowledge of the existence or configuration of a computer system or network. Port Scans or sweeping of a given IP-address range typically fall in this category. (e.g. saint, port sweep, mscan, nmap, etc.).

c) User-to-Root (U2R) attacks: - The aim of gaining root or super-user access on a particular computer or system on which the attacker previously had user level access. These are attempts by a non-privileged user to gain administrative privileges (e.g. Perl, xterm, etc.).

d) Remote-to-Local (R2L) attack :- It is an attack in which a user sends packets to a machine over the internet, which the user does not have access to in order to expose the machine vulnerabilities and exploit privileges which a local user would have on the computer (e.g. xclock, dictionary, guest_password, phf, send mail, xsnoop, etc

2. LITERATURE REVIEW:-

ADAM (Audit Data Analysis and Mining) [1] is an intrusion detector built to detect intrusions using data mining techniques. It first absorbs training data known to be free of attacks. Next, it uses an algorithm to group attacks, unknown behavior, and false alarms. ADAM has several useful capabilities, namely;

*Classifying an item as a known attack
*Classifying an item as a normal event,
*Classifying an item as an unknown attack,
*Match audit trial data to the rules it gives rise to.

IDDM (Intrusion Detection using Data Mining Technique) [2] is a real-time NIDS for misuse and anomaly detection. It applies association rules, met rules, and characteristic rules. It employs data mining to produce description of network data and uses this information for deviation analysis.

MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection) [3] is one of the best known data mining projects in intrusion detection. It is an off-line IDS to produce anomaly and misuse intrusion detection models. Association rules and frequent episodes are applied in MADAM ID to replace hand-coded intrusion patterns and profiles with the learned rules.

In [4], the authors propose a method of intrusion detection using an evolving fuzzy neural network. This type of learning algorithm combines artificial neural network (ANN) and fuzzy Inference systems (FIS), as well as evolutionary algorithms. They create an algorithm that uses fuzzy rules and allow new neurons to be created in order to accomplish this. They use Snort together data for training the algorithm and then compare their technique with that of an augmented neural network.

In [5], a statistical neural network classifier for anomaly detection is developed, which can identify UDP flood attacks. Comparing different neural network classifiers, the back propagation neural network (BPN) has shown to be more efficient in developing IDS . In [6] the author uses the back propagation method by Sample Query and Attribute Query for the Intrusion Detection, analyzing and identifying the most important components of training data. It could reduce processing time, storage requirement, etc.

In [7], Axellson wrote a well-known paper that uses the Bayesian rule of conditional probability to point out that implication of the base-rate fallacy for intrusion detection. In [8], a behavior model is introduced that uses Bayesian techniques to obtain model parameters with maximal a-posterior probabilities. Their work is similar to our, to the extent that Bayesian statistics are employed. However, the difference lies in that; we use naive bayes for our model.

At IBM, Kephart and Arnold [9] developed a statistical method for automatically extracting malicious executable signatures. Their research was based on speech recognition algorithms and was shown to perform almost as good as a human expert at detecting known malicious executables. Their algorithm was eventually packaged with IBM’s antivirus software.

Lo et al. [10] presented a method for filtering malicious code based on “tell-tale signs” for detecting malicious code. These were manually engineered based on observing the characteristics of malicious code. Similarly, filters for detecting properties of malicious executables have been proposed for UNIX systems as well as semiautomatic methods for detecting malicious code

Unfortunately, a new malicious program may not contain any known signatures so traditional signature-base methods may not detect a new malicious executable. In an attempt to solve this problem, the anti-virus industry generates heuristic classifiers by hand. This process can be even more costly than generating signatures, so finding an automatic method to generate classifiers has been the subject of research in the anti-virus community. To solve this problem, different IBM researchers applied Artificial Neural Networks (ANNs) to the problem of detecting boot sector malicious binaries. An ANN is a classifier that models neural networks explored in human cognition. Because of the limitations of the implementation of their Classifier, they were unable to analyze anything other than small boot sector viruses which comprise about 5% of all malicious binary

Using an ANN classifier with all bytes from the boot sector malicious executables as input, IBM researchers were able to identify 80–85% of unknown boot sector malicious executables successfully with a low false positive rate (< 1%). They were unable to find a way to apply ANNs to the other 95% of computer malicious binaries. In similar work, Arnold and Tesauro [11] applied the same techniques to Win32 binaries, but because of limitations of The ANN classifier they were unable to have the comparable accuracy over new Win32 binaries. Our method is different because we analyzed the entire set of malicious executables instead of only boot-sector viruses, or onlyWin32 binaries. Our technique is similar to data mining techniques that have already been applied to Intrusion Detection Systems by Lee et al. [13] . Their methods were applied to system calls and network data to learn how to detect new intrusions. They reported good detection rates as a result of applying data mining to the problem of IDS. We applied a similar framework to the problem of detecting new malicious executables.

3. ANALYSIS OF PROBLEM: -

Now day virus scanner technology has two parts a signature-based detector and a heuristic classifier that detects new viruses. The classic signature-based detection algorithm uses signature of known malicious executables to detect new virus. Signature-based methods create a unique tag for each malicious program so that it can be use as a future examples of it can be correctly classified with a small error rate. These methods do not generalize well to detect new malicious binaries because they are created to give a false positive rate as close to zero as possible. Whenever a detection method generalizes to new instances, the tradeoff is for a higher false positive rate. Heuristic classifiers are generated by a group of virus experts to detect new malicious programs. It is time-consuming and sometime it is not detect new malicious executables.

4. IMPLIMENTED WORK:-

We have used Bayesian classification, in Bayesian classification we have a hypothesis that the given data belongs to particular class. We then calculate the probability for the hypothesis to be true. This is among the most practical approaches for certain types of problems. The approach requires only one scan of the whole data. Also, if at some stage there are additional training data, then each training example can incrementally increase/decrease the probability that a hypothesis is corrected.

4.1 Naive Bayes Classifier :-

The Naïve Bayes model is a heavily simplified Bayesian probability model. In this model, consider the probability of an end result given several related evidence variables. The probability of end result is encoded in the model along with the probability of the evidence variables occurring given that the end result occurs. The probability of an evidence variable given that the end result occurs is assumed to be independent of the probability of other evidence variables given that end results occur.

4.2 KDD Cup Dataset:-

KDD Cup dataset is nothing but Knowledge Discovery dataset, This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition. It is use for building network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

4.3 Classifying Data:-

We are using Kdd cup dataset and Naive Bayes classifier. We classifying the input dataset and detect new Unknown Malicious. Naive Bayes Algorithm is realise on probabilistic condition it seen most probability From the dataset and detect unknown malicious code.

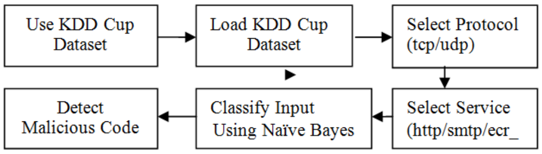
Use KDD Cup Dataset
Load KDD Cup Dataset
Select Protocol (tcp/udp)
Select Service (http/smtp/ecr_
Classify Input Using Naïve Bayes
Detect Malicious Code

Fig. 1 Malicious code Detection using Naive Bayes Classifier.

5. PERFORMANCE ANLYSIS:-

The Naïve bayes classifier have high detection rate and it can be detect malicious code accurately using dataset, it required less CPU cycle, low false positive rate and low false negative rate.

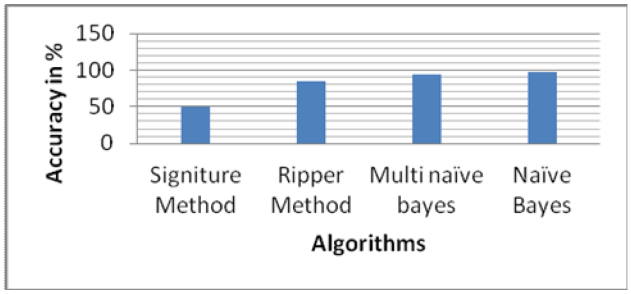

Figure 2 Graph of Accuracy in Percentage

Figure.2 shows the graph of accuracy in percentage. We calculate the accuracy of our project in which we are taken 100 value, insert it and classify it with our dataset it is found that 97 value are found accurate and 3 are found to be Inaccurate from that we calculate that our proposed method naïve bayes has 97% accuracy. Other method like Multi naïve bayes have 95%, Ripper method have83.62% and Signature method have 49.28% accuracy. Our propose method has shown more accurate than other method.

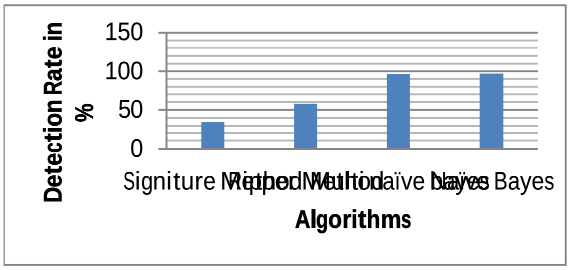

Figure. 3 Graph of Detection rate in percentage

Figure 3 shows the graph of detection rate in percentage. We calculate the detection rate of our project in which we are taken 100 value, insert it and classify it with our dataset it is found that 97 value are found accurately detected from dataset and 3 are found to be inaccurate detected, from that we calculate that our proposed method naïve bayes has 97% detection rate. Other method like Multi naïve bayes have 96%, Ripper method have57.89% and Signature method have 33.75% detection rate. Our proposed method has shown more detection rate than other method.

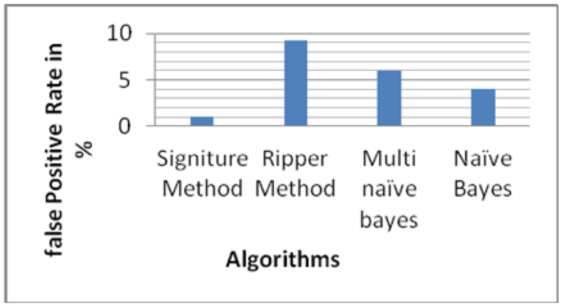

Figure.4 Graph of False positive rate in percentage.

Figure. 4 shows the graph of false positive rate in percentage. False positive rate means a during testing of input data with KDD cup dataset if application shows result data is infectious then may be other possibility is data may not be infectious. Our proposed method naïve bayes have low false positive rate 4%. Other method like Multi naïve bayes has 6%, Ripper method has 9.22% and Signature method have 1% detection rate. Our proposed method has more detection rate and accuracy than other method.

6. APPLICATIONS:-

1. It can be use for detect DOS attacks in networks.
2. It can be use for detect multiple drop packet sources in network.
3. It can be use for detect intrusion in the network for military computers.
4. It can be use for detect intrusion in the network for Big Organization computer.

7. CONCLUSION:-

We have proposed a framework of Network Intrusion Detection System using Naïve Bayes algorithm. The framework classifies the input dataset with KDD cup dataset. The Framework detects attacks in the datasets using the naive Bayes Classifier algorithm. Compared to the neural network based approach, our approach achieve higher detection rate, less time consuming and has low cost factor. However, it generates somewhat more false positives.

References:-

[1] D.Barbara, J.Couto, S.Jajodia, and N.Wu, "ADAM: A test bed for exploring the use of data mining in intrusion detection", SIGMOD, vol30, no.4, pp 15-24, 2001

[2] Tomas Abraham, "IDDM: INTRUSION Detection using Data Mining Techniques", Technical report DSTO electronics and surveillance research laboratory, Salisbury, Australia, May2001.

[3] Wenke Lee and Salvatore J.Stolfo, "A Framework for constructing features and models for intrusion detection systems", ACM transactions on Information and system security (TISSEC), vol.3, Issue 4, Nov 2000.

[4] S.chavan, K.Shah, N.Dave, S.Mukherjee, A. Abraham, and S.Sanyal, "Adaptive neuro-fuzzy Intrusion detection systems", ITCC, Vol 1, 2004

[5] Z. Zhang, J. Li, C.N. Manikapoulos, J.Jorgenson, J.ucles,"HIDE: A hierarchical network intrusion detection system using statistical pre-processing and neural network classification", IEEE workshop proceedings on Information assurance and security, 2001, pp.85-90.

[6] Roy-I Chang, Liang-Bin Lai, et al, "Intrusion detection by back propagation network with sample query and attribute query", International Journal of computational Intelligence Research, Vol.3, no.1, 2007, pp 6-10.

[7] S. Axelsson, "The base rate fallacy and its implications for the difficulty of Intrusion detection", Proc. Of 6th.ACM conference on computer and communication security 1999.

[8] R.Puttini, Z.marrakchi, and L. Me, "Bayesian classification model for Real time intrusion detection", Proc. of 22nd. International workshop on Bayesian inference and maximum entropy methods in science and engineering, 2002.

[9] Jeffrey O. Kephart and William C. Arnold. Automatic Extraction of Computer Virus Signatures. 4thVirus Bulletin International Conference, pages 178-184, 1994

[10] R.W. Lo, K.N. Levitt, and R.A. Olsson. MCF: a Malicious Code Filter. Computers & Security, 14(6):541–566., 1995.

[11] William Arnold and Gerald Tesauro. Automatically Generated Win32 Heuristic Virus Detection. Proceedings of the 2000 International Virus BulletinConference, 2000.

[12] W. Lee, S. J. Stolfo, and P. K. Chan. Learning patterns from UNIX processes execution traces for intrusion detection. AAAI Workshop on AI Approaches to Fraud Detection and Risk Management, pages 50–56. AAAI Press, 1997.

[13] Wenke Lee, Sal Stolfo, and Kui Mok. A Data Mining Framework for Building Intrusion Detection Models. IEEE Symposium on Security and Privacy, 1999

Author:

Aniket.P.Sagane received the B.E. a degrees in Information Technology from Sipna college of Engineering and Technology in 2011 from SGB Amravati University & Pursuing Master of Engineering in Information Technology from Sipna College of Engineering and Technology, from SGBAU Amravati.