Detecting Intrusion in Data Mining using Naive Bayes Algorithm

Shyara Taruna R.[1] Mrs. Saroj Hiranwal[2]
[1]Department of CS&E
[2]Department of Information Technology
[1, 2]SBCET, Jaipur, India

Abstract—With the tremendous increase of network-based services and information sharing on networks, network security is getting more and more importance than ever. Intrusion poses a serious security risk in a network environment. The human classification of the available network audit data instances is usually tedious, time consuming and expensive. Data mining has become a very useful technique for detecting network intrusions by extracting useful knowledge from large number of network data or logs. Naïve Bayes classifier is one of the most popular data mining algorithms for classification, which provides an optimal way to predict the class of an unknown example. We tested the performance of our proposed algorithm by employing KDD99 benchmark network intrusion detection dataset, and the experimental results proved that it improves detection rates as well as reduces false positives for different types of network intrusions.

Key words: Data Mining, Detection Rate, Falser Positive, Intrusion Detection, Naïve Bayes Classifier, Network Security.

I. INTRODUCTION

With the tremendous growth of network-based services and sensitive information on networks, network security is becoming more and more importance than ever before. Intrusion detection techniques are the last line of defenses against computer attacks behind secure network architecture design, firewalls, and personal screening.

Intrusion detection systems (IDSs) play a very important role in network security. Intrusion detection systems (IDSs) is security tools that collect information from a variety of network sources, and analyze the information for signs of network intrusions. IDS can be host-based or network-based systems [1]. Host-based IDS locates in servers to examine the internal interfaces, and network-based IDS monitors network packets to discover network intrusions. The success of an IDS can be characterized in both detection rates (DR) and false positives (FP) for different types of intrusions [2].

This paper presents the scope and status of our research in anomaly detection. This paper gives a comparative study of several anomaly detection schemes for identifying novel network intrusion detections. We present experimental results on KDDCup'99 data set. Experimental results have demonstrated that our naïve bayes classifier model is much more efficient in the detection of network intrusions, compared to the neural network based classification techniques.

II. INTRUSION DETECTION

An Intrusion Detection System (IDS) inspects the activities in a system for suspicious behavior or patterns that may indicate system attack or misuse.

An IDS monitors network traffic in a computer network like a network sniffer and collects network logs. Then the collected network logs are analyzed for rule violations by data mining algorithms. When any rule violation is detected, the IDS alert the network security administrator or automated intrusion prevention system (IPS). Intrusion detection system can be classified into three systems based on such (i) misuse based system, (ii) anomaly based systems, and (iii) hybrid systems [3] – [8]. Misuse based IDS simple pattern matching techniques to match the attack pattern, and a database of known attack patterns are consistent, and produce very low false positive (FP). It requires the signature of the rules or to see, not so well-known attacks regularly updated. Anomaly based of the IDS to determine the normal behavior by examining the abnormal behavior of the new attack [9], both well-known and achieve a high detection rate (DR) unknown attacks, but makes many false positives (FP). Anomaly based IDS, the development of IDS audit data collected by observing the rules. Developed by the operating system audit data record of the activities is logged to a file in chronological order. On the other hand, a combination of a hybrid IDS based on misuse and corruption of the detection system technology. The current adaptive intrusion detection is designed to address large amounts of data in the analysis of audit, inspection rules for performance optimization.

III. NETWORK ATTACKS

The simulated attacks were classified, according to the actions and goals of the attacker. Each attack type falls into one of the following four main categories [10]:

Denials-of Service (DoS) attacks have the goal of limiting or 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, mailbomb, udpstorm, SYNflood, etc.).

Probing or Surveillance attacks have the goal of gaining knowledge of the existence or configuration of a computer system or network. Port Scans or sweeping of a given IP-address range typ ically fall in this category. (e.g. saint, portsweep, mscan, nmap, etc.).

User-to-Root (U2R) attacks have the goal 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.).

Remote-to-Local(R2L) attack 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, sendmail, xsnoop, etc.).

IV. THE PROPOSED METHOD

The Naïve Bayes method is based on the work of Thomas Bayes (1702-1761). In Bayesian classification, we have a hypothesis that the given data belongs to a 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 correct. Thus, a Bayesian network is used to model a domain containing uncertainty [11] – [12].

Consider the following example where a farmer has a bottle of milk that can be either infected or clean. She also has a test that determines with a high probability whether the milk is infected or not (i.e. the outcome of the test is either positive or negative). This situation can be represented with two random variables, infected and positive. The variable infected is true when the milk is actually infected and false otherwise. The variable positive is true when the test claims that the milk is infected and false when the outcome of the test is negative. Note that, it is possible that the milk is clean when the test data has a positive outcome and vice-versa.

Procedure: Decision Tree
Input: Dataset D
Output: DA, FP For Attack Data

Do
Take the Class CL From D.
For each attribute value
Remove the noise from the dataset.
Calculate the prior probability $P(C_j)$ for each class C_j in dataset.

$$D: P(C_j) = \frac{\sum t_i \rightarrow c_j}{\sum_{i=1}^{n} t_i}$$

End For
For each attribute value
Calculate the class conditional probabilities $P(A_{ij}|C_j)$ for each attribute values in dataset D:

$$P(A_{ij} \mid C_j) = \frac{\sum_{i=1}^{n} A_i \rightarrow c_j}{\sum t_i \rightarrow c_i}$$

End For
End Do
Do
Multiply the prior probability and class conditional probability.
End Do
Do
Consider the class with the highest classifier probability.
End Do
Repeat steps 2to4 until all attribute at their highest probability

In this Algorithm first we find out the prior probability for the give intrusion data set then find out the class conditional probability for the data set. After that we find out the highest classifier probability and base on we find out the Detection Rate and false positive for the intrusion data set. To find out the Gain ratio first find out the Gain for all the attribute for the data set then find out Split Info for each and every attribute so this way we find out gain ratio for the intrusion data set. The prior probability P (C_j) for each class is estimated by counting how often each class occurs in the dataset Di. For each attribute Ai the number of occurrences of each attribute value A_{ij} can be counted to determine $P(A_i)$. The class conditional probability $P(A_{ij}|C_j)$ for each attributes values A_{ij} can be estimated by counting how often each attribute value occurs in the class in the dataset D.

The naïve Bayes model is a heavily simplified Bayesian probability model [13]. 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. Now we will consider the alarm example using a naïve Bayes classifier. Assume that we have a set of examples that monitor some attributes such as whether it is raining, whether an earthquake has occurred etc. Let's assume that we also know, using the monitor, about the behaviour of the alarm under these conditions. In addition, having knowledge of these attributes, we record whether or not a theft actually occurred. We will consider the category of whether a theft occurred or not as the class for the naïve Bayes classifier. This is the knowledge that we are interested in. The other attributes will be considered as knowledge that may give us evidence that the theft has occurred. Figure1 below shows the framework for a Naïve Bayesian model to perform intrusion detection.

The naïve Bayes classifier operates on a strong independence assumption [14]. This means that the probability of one attribute does not affect the probability of the other. Given a series of n attributes, the naïve Bayes classifier makes 2n! Independent assumptions. Nevertheless, the results of the naïve Bayes classifier are often correct. The work reported in [15] examines the circumstances under which the naïve bays classifier performs well and why. It states that the error is a result of three factors: training data noise, bias, and variance. Training data noise can only be minimized by choosing good training data. The training data must be divided into various groups by the machine learning algorithm. Bias is the error due to groupings in the training data being very large.

Variance is the error due to those groupings being too small.

V. INTRUSION DETECTION DATASET

The KDD99 cup dataset was used in the 3rd International Knowledge Discovery and Data Mining Tools Competition for building a network intrusion detector, a predictive model capable of distinguishing between intrusions and normal connections [16]. In 1998, DARPA intrusion detection evaluation program, a simulated environment was set up to acquire raw TCP/IP dump data for a local-area network (LAN) by the MIT Lincoln Lab to compare the performance of various intrusion detection methods. It was operated like a real environment, but being blasted with multiple intrusion attacks and received much attention in the research community of adaptive intrusion detection. The KDD99 dataset contest uses a version of DARPA98 dataset. In KDD99 dataset, each example represents attribute values of a class in the network data flow, and each class is labeled either normal or attack. The classes in KDD99 dataset can be categorized into five main classes (one normal class and four main intrusion classes: probe, DOS, U2R, and R2L).

In KDD99 dataset these four attack classes (DoS, U2R, R2L, and probe) are divided into 22 different attack classes that tabulated in Table I.

4MainAttack Classes	**22 Attack Classes**
Denial of Service (DoS)	back, land, neptune, pod, smurt, teardrop
Remote to User (R2L)	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
User to Root (U2R)	buffer_overflow, perl, loadmodule, rootkit
Probing	ipsweep, nmap, portsweep, satan

Table. 1: Attack Classes In KDD99 Dataset

There are total 41 input attributes in KDD99 dataset for each network connection that have either discrete or continuous values and divided into three groups. The first group of attributes is the basic features of network connection, which include the duration, prototype, service, number of bytes from source IP addresses or from destination IP addresses, and some flags in TCP connections. The second group of attributes in KDD99 is composed of the content features of network connections and the third group is composed of the statistical features that are computed either by a time window or a window of certain kind of connections. Table II shows the number of examples of 10% training data and 10% testing data in KDD99 dataset. There are some new attack examples in testing data, which is no present in the training data.

Attack Types	Training Examples	Testing Examples
Normal	97277	60592
Denial of Service	391458	237594
Remote to User	1126	8606
User to Root	52	70
Probing	4107	4166
Total Examples	494020	311028

Table. 2: Number of Examples in Training and Testing KDD99 Data

VI. CONCLUSION

This paper introduced a learning algorithm for detecting network intrusions using naive Bayesian classifier with data mining. The algorithm is suitable for analyzing large number of network logs or audit data. It improves the performance of detection rates for different types of intrusions. The main propose of this paper is to improve the performance of naïve Bayesian classifier for intrusion detection. We tested out proposed algorithm on KDD99 dataset that shows it maximized the balance detection rates for 4 attack classes in KDD99 dataset and minimized false positives at acceptable level. The future work focus on apply this algorithm in real time network and ensemble with other data mining algorithms for improving the detection rates in intrusion detection.

REFERENCES

[1] "Symantec-Internet Security threat report highlights Symantec.com)", http://www.prdomain.com/companies/Symantec/newreleases/Symantec_internet_205032.htm

[2] R. Durst, T. champion, B. Witten, E. Miller, and L. Spagnuolo, "Testing and evaluating computer intrusion detection system" communications of ACM, Vol.42, no.7, pp 53-61, 1999.

[3] Taruna R. Shyara & Priyanka Trikha, "A Framework: Intrusion Detection in Data Mining" International Journal of Research in Computer Engineering and Electronics. ISSN 2319-376X Vol: 2 ISSUE: 3, June-2013.

[4] D. Barbara, J. Couto, S. Jajodia, L. Popyack and N. Wu, "ADAM: Detecting intrusion by data mining," IEEE Workshop on Information Assurance and Security, West Point, New York, June 5-6, 2001.

[5] W. Lee, "A data mining and CIDF based approach for detecting novel and distributed intrusions," Recent Advances in Intrusion Detection, 3rd International Workshop, RAID 2000, Toulouse, France, October 2-4, 2000, Proc. Lecture Notes in Computer Science 1907 Springer, 2000.

[6] R. Wasniowski, "Multi-sensor agent-based intrusion detection system," In Proc. of the 2nd Annual Conference on Information Security, Kennesaw, Georgia, 2005, pp. 100-103.

[7] N.B. Amor, S. Benferhat, and Z. Elouedi, "Naïve Bayes vs. decision trees in intrusion detection systems," In Proc. of 2004 ACM Symposium on Applied Computing, 2004, pp. 420-424.

[8] W. Lee, S. J. Stolfo, and K. W. Mok, "Adaptive Intrusion Detection: A Data Mining Approach," Artificial Intelligence Review, 14(6), December 2000, pp. 533-567.

[9] Lazarevic, A., Ertoz, L., Kumar, V., Ozgur,. A., Srivastava, and J., "A comparative study of anomaly detection schemes in network intrusion detection," In Proc. of the SIAM Conference on Data Mining, 2003.

[10] A.Sung & S.Mukkamala, "Identifying important features for intrusion detection using SVM and neural networks," in symposium on application and the Internet, pp 209-216,2003.
[11] P.Jenson, "Bayesian networks and decision graphs", Springer, New-york, USA, 2001.
[12] J.Pearl, "Probabilistic reasoning in intelligent system", Networks of plausible inference, Morgan Kaufmann 1997.
[13] S.J.Russel, and Norvig, "Artificial Intelligence: A modern approach (International edition), Pearson US imports & PHIPES, Nov 2002.
[14] P.Domingos, and M.J. Pizzani, "On the optimality of the simple Bayesian classifier under zero-one loss", m/c learning, Vol.29, no2-3, pp 103-130, 1997.
[15] M.Mahoney and P. chan, "An analysis of the 1999 DARPA/Lincoln laboratory evaluation data for network anomaly detection", Proc.of Recent Advances in intrusion detection (RAID)-2003, Pittsburg, USA, Sept. 2003.
[16] The KDD Archive.KDD99 cup dataset,1999 .http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

