

Enhanced Naïve Bayes Algorithm for Intrusion Detection in Data Mining

Shyara Taruna R.¹ Mrs. Saroj Hiranwal²

¹Department of CS & E, SBTC, Jaipur, India

²Department of Information Technology, SBTC, Jaipur, India

Abstract -Classification is a classic data mining technique based on machine learning. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. Naïve Bayes is a commonly used classification supervised learning method to predict class probability of belonging. This paper proposes a new method of Naïve Bayes Algorithm in which we tried to find effective detection rate and false positive rate of given data. 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.

Keywords:

Data Mining, Detection Rate, False Positive, Intrusion Detection, Naïve Bayes Classifier, Network Security.

I. INTRODUCTION

Data Mining [1] is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management System.

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 [2]. 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 [3].

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 [4] – [9]. 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 [10], 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 [11]:

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 typically 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 Enhanced Naïve Bayes method is based on the work of Thomas Bayes (1702-1761) and Naïve Bayes Algorithm for Intrusion Detection. In this algorithm first we find out the prior probability for the given intrusion data set then find out class conditional probability for the data set. After that we find out the highest classifier probability and base on we find detection rate and false positive 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 D_i . For each attribute A_i the number of occurrences of each attribute value A_{ij} can be determine $P(A_i)$. The class conditional probability $P(A_{ij}|C_j)$ for each attribute values A_{ij} can be estimated by counting how often each attribute value occurs in the class in the dataset D .

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.

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

End For

For each attribute value

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

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

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 D_i . For each attribute A_i 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 .

V. INTRUSION DETCTION 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.

Attack Type	Attack Name
Normal	Normal
DOS	Smurf, Neptune, Back, Teardrop, Pod Land
U2R	Buffer, Overflow, Rootkit, Load Module, Perl
R2L	Warezcilent, Guess_passwd, Warezmaster, Imap, ftp_write, Multihop, Phf, Spy
PROBE	Satan, Ipsweep, Portsweep, Nmap

TABLE I. 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.

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 purpose 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.

VII. 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., Ertöz, 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>

AUTHOR

1. **Shyara Taruna R.** is pursuing M.Tech. in Computer Science and Engineering from Rajasthan Technical University, India. Her research is focused on Intrusion Detection in Data Mining.

2. **Mrs. Saroj Hiranwal** is co-author of paper. She Completed her M.Tech. in Information Technology in 2006. She is pursuing her Ph.D. Currently she is Head of IT Department in Sri Balaji Technical Campus, Jaipur, Rajasthan. She guide many M.Tech. students for dissertation.