An Efficient Machine Learning and Data Mining Method for Finding Anomalies in a Cyber Security Intrusion Detection System

Marpu Gowtami[1], Mula Sudhakar[2]

Final M.Tech Student[1], Asst.professor[2]
[1,2]Dept of CSE, Sarada Institute of Science, Technology and Management (SISTAM), Srikakulam, Andhra Pradesh

Abstract:
Now a day's network security is one of the most important concerns in modern era. With the rapid development of technology and most usage of internet will increase daily. So that one of the vulnerability is network security have become important issue in the network. Intrusion detection system is used to identify unauthorized users and also unusual attacks over the secured networks. Over the past years, many studies have been conducted on the intrusion detection system. However, in order to understand the current status of implementation of machine learning techniques for solving the intrusion detection problems. An Intrusion Detection System (IDS) is designed to detect system attacks and classify system activities into normal and abnormal form. Machine learning techniques have been applied to intrusion detection systems which have an important role in detecting Intrusions. In this paper we are implementing classifier algorithms for finding unauthorized users and also overcome attacks on secured networks. This paper also presents the system design of an Intrusion detection system to reduce false alarm rate and improve accuracy to detect intrusion.

Keywords: *intrusion detection, classification, Anomaly, Prior Probability.*

I. Introduction

Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms [1]. The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type of attack [2][3]. We are also interested in link and sequence analysis [4][5][6]. Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the security

analyst in identifying areas of concern [7]. The models must be represented in some form. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions (including neural nets), instance-based examples, and probability models [1].

Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests. In contrast to statistical techniques, machine learning techniques are well suited to learning patterns with no a priori knowledge of what those patterns may be. Clustering and Classification are probably the two most popular machine learning problems. Techniques that address both of these problems have been applied to IDSs.

The idea of applying machine learning techniques for intrusion detection is to automatically build the model based on the training data set. This data set contains a collection of data instances each of which can be described using a set of attributes (features) and the associated labels. The attributes can be of different types such as categorical or continuous. The nature of attributes determines the applicability of anomaly detection techniques. For example, distance-based methods are initially built to work with continuous features and usually do not provide satisfactory results on categorical attributes. The labels associated with data instances are usually in the form of binary values i.e. normal and anomalous. In contrast, some researchers have employed different types of attacks such as DoS, U2R, R2L and Probe rather than the anomalous label. This way learning techniques is able to provide more information about the types of anomalies. However, experimental results show that current learning techniques are not precise enough to recognize the type of anomalies. Since labeling is often done manually by human experts, obtaining an accurate labeled data set which is representative of all types of behaviors is quite expensive. As a result, based on the availability of the labels, three operating modes are defined for anomaly detection techniques: as Supervised Learning, Unsupervised Learning, Semi supervised Learning.

Classification Techniques: In a classification task in machine learning, the task is to take each instance of a dataset and assign it to a particular class. A classification based IDS attempts to classify all traffic as either normal or malicious. The challenge in this is to minimize the number of false positives (classification of normal traffic as malicious) and false negatives (classification of malicious traffic as normal). A framework of NIDS based on a Naïve Bayes algorithm is proposed in [19]. The framework constructs the patterns of the network services over data sets labeled by the services. The framework detects attacks in the datasets using the naïve Bayes Classifier algorithm using the built patterns. Compared to the Neural network based approach, their approach achieves higher detection rate, less time consuming and has a low cost factor. However, it generates more false positives. Naive Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes. This poses a limitation of this research work. In order to minimize this problem so as to reduce the false positives, active platform or event based classification may be thought of using Bayesian network.

II. RELATED WORK

One way to improve certain properties, such as accuracy, of a data mining system is to use a multiplicity of techniques and correlate the results together. The combined use of numerous data mining methods is known as an ensemble approach, and the process of learning the correlation between these ensemble techniques is known by names such as multistrategy learning, or meta-learning.

Researchers at the Columbia University IDS lab have applied meta-classification both to improve accuracy and efficiency, as well as to make data mining based IDS systems more adaptable. Lee and Stolfo [5] used meta-classification to improve both accuracy and efficiency (by running high cost classifiers only when necessary), and combined the results using boolean logic. The classifiers are produced using cost factors that quantify the expense of examining any particular feature in terms of processing time, versus the cost of responding to an alert or missing an intrusion [8][9]. Didaci et al. [10] applied a meta-classification approach. The authors applied three different classification methods - the majority voting rule, the average rule, and the ”belief” function - to the outputs of three distinct neural nets. The Neural nets had previously been trained on different features sets from the KDD tcpdump data. They found that these multistrategy techniques, particularly the belief function, performed better than all three neural nets individually. Crosbie and Spafford [11] use an ensemble of ”autonomous agents” to determine the threat level presented by network activity.

Based on the observation that an intrusion scenario might be represented as a planning activity, Cuppens et al [12] suggest a model to recognize intrusion scenarios and malicious intentions. This model does not follow previous proposals that require to explicitly specify a library of intrusion scenarios. Instead, their approach is based on specification of elementary attacks and intrusion objectives. They then show how to derive correlation relations between two attacks or between an attack and an intrusion objective. Detection of complex intrusion scenario is obtained by combining these binary correlation relations. They also suggest using abduction to recognize intrusion scenarios when some steps in these scenarios are not detected. They then define the notion of anti-correlation that is useful to recognize a sequence of correlated attacks that does no longer enable the intruder to achieve an intrusion objective. This may be used to eliminate a category of false positives that correspond to false attacks.

III.PROPOSED SYSTEM

One of the main challenges in the security management of large-scale high-speed networks is the detection of suspicious anomalies in network. So that to identify anomalies in the network we are using the machine learning technologies in a data mining. By implementing machine learning technologies we are using classification technique in the data mining. Before performing the classification process we are take the two types of data sets. By using those data sets we are identifying anomalies or not in the network. In this paper we are take net flow data variables and that data should contain unidirectional sequence of packets. Net Flow data include a compressed and pre-processed version of the actual network packets. The statistics are derived features and, based on certain parameters such as duration of window, number of packets, etc., set the Net Flow settings on the device. The net flow performs data streaming process, refinement analysis and classification. The net flow continuous captured data in the online stream analysis. The data captured by net flow framework with contains attributes such as source address, destination address, source port, destination port, type of protocol and packet size.

In general anomaly detection can be considered as reactive because when system responds as input when input is unexpected. Conversely, in a misuse detection problem, the system is proactive because the signatures extracted from the input are being checked continuously against a list of attack patterns. Taking the proactive approach as in misuse detection requires classifying the network streams. In their work we are proposed a naïve Bayesian classifier technique for identifying anomalies or not. Before performing the classification approach we can take the two types data set with containing above specified attributes. The first data set name as training data with containing above attributes along the status of each source node to destination. Another one data set testing data with contain only above specified attributes. By taking those two data sets we can apply the naïve bayeseian classifier technique we can specify the status of each node in a testing data. The implementation of procedure of this paper is as follows.

Reading the Training Data Set:

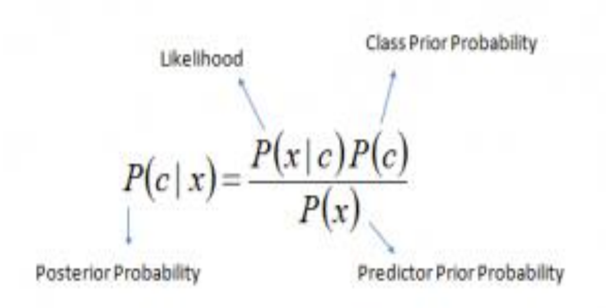
In this module we are retrieve the training data set from the database. By using this training data set we are identify anomaly detection stage and also specify normal stage or attack decision is made by using this data set. In the training data set contain attributes related to source address, destination address, source port number, destination port number, type of protocol and size of packet. Based those attribute we can already contains each node status for normal user and anomaly user. By taking those attribute value we can identify of run user status. In this paper we are identify more than one user status by taking testing data. The testing data will take and pass this testing through training data set by using classification approach; we can get status related for testing data members.

Classification process using Naïve Bayesian Technique:

In this module we are taking two data sets like training and testing, apply the classification process for finding normal or anomalies users. By performing classification process we are using naïve Bayesian classification approach. It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



$$P(c\,|\,X) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

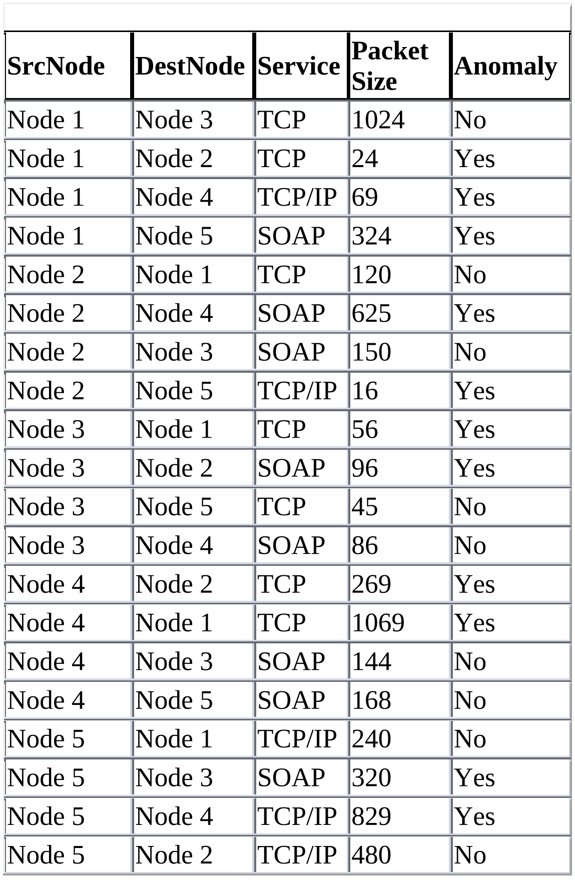
Above,

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Let us understand above algorithm we taking an example of training data set with contain attributes like source node address, destination node address, source port number, destination port number, type of protocol and packet size. Now we need to classify the given node is normal node or anomaly node based on considering training data set attributes. Lets follow the below steps to perform the classification approach.

Step1. Convert data set in to frequency table.

Step 2. Create like hood table by finding the probabilities by taking below table data.



SrcNode	DestNode	Service	Packet Size	Anomaly
Node 1	Node 3	TCP	1024	No
Node 1	Node 2	TCP	24	Yes
Node 1	Node 4	TCP/IP	69	Yes
Node 1	Node 5	SOAP	324	Yes
Node 2	Node 1	TCP	120	No
Node 2	Node 4	SOAP	625	Yes
Node 2	Node 3	SOAP	150	No
Node 2	Node 5	TCP/IP	16	Yes
Node 3	Node 1	TCP	56	Yes
Node 3	Node 2	SOAP	96	Yes
Node 3	Node 5	TCP	45	No
Node 3	Node 4	SOAP	86	No
Node 4	Node 2	TCP	269	Yes
Node 4	Node 1	TCP	1069	Yes
Node 4	Node 3	SOAP	144	No
Node 4	Node 5	SOAP	168	No
Node 5	Node 1	TCP/IP	240	No
Node 5	Node 3	SOAP	320	Yes
Node 5	Node 4	TCP/IP	829	Yes
Node 5	Node 2	TCP/IP	480	No

By using the above table we can calculate probabilities based like hood table.

Anomaly	Yes	No	
Node1	3	1	3/20 = 0.15

Step3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Problem: node1 is not anomaly. Is this statement is correct?
We can solve it using above discussed method of posterior probability.

P(Yes | Node1) = P(Node1 | Yes) * P(Yes) / P (Node1)

Here we have P (Node1 |Yes) = 3/20 = 0.15, P(Node1) = 4/20 = 0.2, P(Yes)= 11/20 = 0.55

Now, P (Yes | Node1) = 0.15 * 0.2 / 0.55 = 0.05, which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes

IV. CONCLUSIONS

In this paper we are proposed an efficient machine learning technology for identifying normal or anomaly users in the network. By implementing anomaly intrusion process we are using classification technique in the data mining. In this paper we are implementing naïve Bayesian algorithm is used for intrusion detection in cyber security system. Before applying this technique we are taking training data set contains information related of all nodes with anomaly status. We are taking another data set is testing data set for information contains running status of all nodes in network. By taking this data set and apply classification approach on the training data set, we can identify normal or anomaly users in the network. So that by implementing naïve Bayesian algorithm we can get efficient result and also improve performance by taking a large testing data.

REFERENCES

[1]. Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth, ”The KDD process for extracting useful knowledge from volumes of data,” Communications of the ACM 39 (11), November 1996, 2734.

[2]. Ghosh, A. K., A. Schwartzbard, and M. Schatz,”Learning program behavior profiles for intrusion detection”, In Proc. 1st USENIX, 9-12 April, 1999

[3]. Kumar, S., ”Classification and Detection of Computer Intrusion”, PhD. thesis, 1995, Purdue Univ., West Lafayette, IN.

[4]. Lee, W. and S. J. Stolfo, ”Data mining approaches for intrusion detection”, In Proc. of the 7th USENIX Security Symp., San Antonio, TX. USENIX, 1998.

[5]. W. Lee, S.J.Stolfo et al, ”A data mining and CIDF based approach for detecting novel and distributed intrusions”, Proc. of Third International Workshop on Recent Advancesin Intrusion Detection (RAID 2000), Toulouse, France.

[6] Lee, W., S. J. Stolfo, and K. W. Mok, ”A data mining framework for building intrusion detection models,” In Proc. of the 1999 IEEE Symp. On Security and Privacy, Oakland, CA, pp. 120132. IEEE Computer Society Press, 9-12 May 1999

[7.] Eric Bloedorn et al,”Data Mining for Network Intrusion Detection: How to Get Started,” Technical paper, 2001.

[8]. Fan, W., W. Lee, S. J. Stolfo, and M. Miller, ”A multiple model cost sensitive approach for intrusion detection”, In R. L. de M’antaras and E. Plaza (Eds.), Proc. of Machine Learning: ECML 2000, 11th European Conference on Machine Learning, Volume 1810 of Lecture Notes in Computer Science, Barcelona, Spain, pp. 142153. Springer, 31 May - 2 June, 2000.

[9]. Fan, W., ”Cost-Sensitive, Scalable and Adaptive Learning Using Ensemble-based Methods”, Ph. D. thesis, Columbia Univ., 2001.

[10]. Didaci, L., G. Giacinto, and F. Roli, ”Ensemble learning for intrusion detection in computer networks”, Proc. of AI*IA,

Workshop on ”Apprendimento automatico: metodi e applicazioni”, Sept 11, 2002, Siena, Italy.

[11]. Crosbie, M. and E. H. Spafford, ”Active defense of a computer system using autonomous agents”, Technical Report CSD-TR- 95-008, Purdue Univ., West Lafayette, IN, 15 February 1995.

[12]. F. Cuppens, F. Autrel, A. Miege, and S. Benferhat, ”Correlation in an intrusion detection process”, In S Ecurit e des Communications sur Internet (SECI’02), Sep. 2002.

[13]. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, “Anomaly-based network intrusion detection: Techniques, systems and challenges,” *Comput. Secur.*, vol. 28, no. 1, pp. 18–28, 2009.

BIOGRAPHIES:

Marpu Gowtami is student in M.Tech(CSE) in Sarada Institute of Science Technology and management, srikakulam. She has received her B. Tech (IT) from Prajna Iinstitute of Technology and management, Ramakrishnapuram, palasa. Her interesting areas are data mining, network security and cloud computing.

Mula Sudhakar is working as an Assistant Professor in Sarada Institute of Science, Technology and Management, Srikakulam, Andhra Pradesh. He received his M.Tech (SE) from Sarada Institute of Science, Technology and Management, Srikakulam. Andhra Pradesh. His research areas include Computer Networks, Data Mining, and Distributed Systems.