Network Intrusion Detection Using Machine Learning

Md Nasimuzzaman Chowdhury and Ken Ferens, Mike Ferens[1]
Department of Electrical and Computer Engineering University of Manitoba
Winnipeg, Manitoba, Canada
[1]Gourdie-Fraser, Inc., Traverse City, Michigan, US
{chowdhmn@myumanitoba.ca, ken.ferens@umanitoba.ca}

Abstract— *In the network communications, network intrusion is the most important concern nowadays. The increasing occurrence of network attacks is a devastating problem for network services. Various research works are already conducted to find an effective and efficient solution to prevent intrusion in the network in order to ensure network security and privacy. Machine learning is an effective analysis tool to detect any anomalous events occurred in the network traffic flow. In this paper, a combination of two machine learning algorithms is proposed to classify any anomalous behavior in the network traffic. The overall efficiency of the proposed method is dignified by evaluating the detection accuracy, false positive rate, false negative rate and time taken to detect the intrusion. The proposed method demonstrates the effectiveness of the algorithm in detecting the intrusion with higher detection accuracy of 98.76% and lower false positive rate of 0.09% and false negative rate of 1.15%, whereas the normal SVM based scheme achieved a detection accuracy of 88.03% and false positive rate of 4.2% and false negative rate of 7.77%.*

Keywords—Intrusion Detection; Machine Learning; Support Vector Machine, Supervised Learning

1. Introduction

Network Security maintenance is one of the major safety concerns for neutralizing any unwanted activities. It is not only for protecting data and network privacy issues but also for avoiding any hazardous situations. From January through June 2010 Microsoft security intelligence report shows that the infection trends are still increasing on average around the world at a higher rate [1]. For decades, Network security is one of the major issues and different types of developed systems are being implemented. Network intrusion is an unauthorized activity over the network that steals any important and classified data. Also sometimes it's the reason of unavailability of network services. The unexpected anomaly occurs frequently and a great loss to internet cyber world in terms of data security, the safety of potential information's etc. Therefore, the security system has to be robust, dependable and well configured. Principally it is of two types on network intrusion detection. One is signature based and another is anomaly based detection system. Signature based detection system involves analyzing network traffic for a series of bytes or packet sequences known to be an anomaly. A major disadvantage of this detection scheme is that signatures are comparatively fair easier to develop and understand if one knows what network behavior need to be identified. Signature based type detection also has some disadvantages. A signature needs to be created for each attack and they are able to detect only those attacks. They are unable to detect any other novel attacks as their signatures are unknown to the detection scheme.

The Anomaly based type detection scheme concept is based on analyzing the characteristic of the network behavior. This type detection has the capability to detect anomaly behavior by analyzing the high volume traffic, a surge in traffic from a specific host or to a specific host, load imbalance in the network [2]. One disadvantage of this kind of scheme is that if the malicious behavior falls within normal network behavior then it's not detected as an anomaly. Major Advantages over the signature based is, a new attack for which a signature does not exist can be detected if it behaves differently from normal traffic behavior patterns. For data confidentiality, classified data security and for preventing unauthorized access detection of intrusion is an essential task for ensuring the security of the networks.

There are several types of method proposed for network intrusion detection. The anomaly network intrusion detection is a major part of network security [3], [4]. Sometimes the behavior of the anomaly seems to be similar as normal data usage [5]. One problem in anomaly detection refers to the issue of classification problem that how to make a distinction between normal and abnormal activities in an effective and efficient way.

Presently machine learning system has been extended for implementing effective intrusion detection system. Machine learning methods are very functional and improved in current intrusion detection. In particular, support vector machines [6], neural networks [7], decision trees seems to have efficient significant schemes in anomaly detection systems to improve the classification performance and speed.

In this paper, an new algorithm is proposed using a combination of two machine learning methods Simulated Annealing & Support Vector Machine that can detect any anomalous behavior of the network and can able to classify between normal and abnormal behavior. It doesn't require any hardware specifications and can be used for pattern recognition of the malicious behavior.

2. Support Vector Machine

Support vector machines (SVM) [8], a type of machine learning method; capable of being performing a range of classification tasks. It's also a set of related supervised learning methods that can analyze data and recognize patterns.

SVMs have been evolved to give a standard generalized performance to solve wide range classification and pattern recognition problems such as handwritten character recognition [9], face detection [10], pedestrian detection [11], and text categorization.

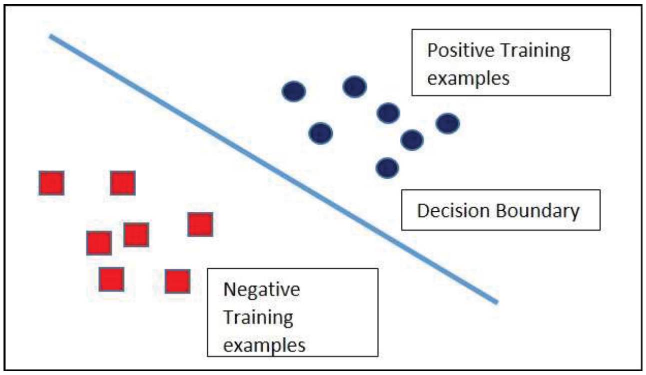

Figure 1: A linear Support Vector Machine.

Considering a training dataset $\{X_i , Y_i\}_{i=1}^{n}$, where X_i represents the input vector of svm that contains the n dimensional input features and $Y_i \in \{+1, -1\}$ represents the output. $Y_i = 1$ Denotes the positive group of training samples and $Y_i = -1$ denotes the negative training samples.

The decision surface in the form of hyperplane is defined as

$$W.X + b = 0 \quad (1)$$

Where,

W= Weight Vector
b = The bias

Linear SVM maximizes the geometric margin of training dataset.

$$max_{w,b}\ C$$

s.t

$$Y_i \frac{1}{||w||}(W.X_i + b) \geq C, i = 1. \dots n \quad (2)$$

Where C is called the regularization parameter.

Any solution that can be found within the constrain boundary, any positively scaled multiples will satisfy them too. So if

$||w|| = 1/C$, the linear SVM can be formulated as

$$\min||x|| \leftrightarrow \frac{\min\ \ 1}{2\ ||x||^2}$$

s.t

$$Y_i(W.X_i + b) \geq 1, i = 1, \dots \dots . . n \quad (3)$$

With this settings a margin around the linear decision boundary can be shown at a higher dimension

3. Selected Dataset and Feature Description

The dataset used for this research paper were chosen from the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS). In this dataset, a hybrid of real modern normal activities and attack behaviors were generated. This dataset contains total forty-seven features and also contains over 2 million sample data [12] [13].

Feature selection is the most important step for network intrusion detection. Features play an important role to achieve classification accuracy which improves the effectiveness and also the efficiency of an intrusion detection system. In this paperwork, to detect an anomaly and to achieve the highest detection accuracy with the shortest possible time, a set of three features were randomly selected each time by the algorithm. Then SVM was performed among those randomly selected features to find the detection accuracy, false positive, false negative and time taken for anomaly detection.

The following table describes a few randomly selected features combinations that have been used for detecting anomalous and normal behavior in the network data traffic [12] [13].

Table 1: Features

Combination Number	**Features in this Combination**
1	a. Source transmission control base sequence number **b**. Source TCP window advertisement value **c**. Source and Destination IP address equal and port numbers
2	**a.** If the ftp session is time series is accessed by user and password **b**. Number of flows that has a command in ftp session. **c**. Number of connections in same source and destination address in past 100 connections
3	**a.** Source transmission control connection setup round-trip time **b**. Source and Destination IP address equal and port numbers **c**. No. of connections in same destination address and source port in past 100 connections
4	**a.** Destination TCP base sequence number **b**. Source and Destination IP address equal and port numbers **c**. No. of connections in same destination address and source port in past 100 connections
5	**a.** From the Source to destination time to live value while the packets are alive **b**. Source TCP window advertisement value **c.** No. of connections that contain same service and source address in previous 100 connections
6	**a.** If the ftp session is time series is accessed by user and password **b**. No. of connections in same source address and destination port **c**. No. of connections in same source and destination address in past 100 connections
7	**a.** Record start time **b**. Source inter packet arrival time (mSec) **c**. No. of connections that contain same service and source address
8	**a** From the Source to destination time to live value while the packets are alive **b**. Destination to source packet count **c**. No. of connections in same destination address and source port in past 100 connections
9	**a.** From the Source to destination time to live value while the packets are alive **b.** Destination inter packet arrival time (mSec) **c**. No. of connections in same source address
10	**a.** Destination packets retransmitted or dropped **b**. Destination to source packet count **c**. Source jitter (mSec)

4. Proposed Machine Learning Algorithm

The proposed algorithm is based firstly on simulated annealing that makes random combinations of 3 features at a time and then SVM is applied on that feature combination that is able to detect anomalous behavior from the internet data traffic. The details of the proposed algorithm is given below:

1. Define the number of features, K from the dataset.

2. Select n features among K features using random combination where $n \in K$
3. Run SVM on n featured training examples
   a) Select the total number of N data samples (n featured) to run the SVM.
   b) Select SVM parameter (Gamma, coef θ, nu etc.)
   c) Select $n \times N$ data sample for training and save the data on T_{train} dataset
   d) Select $n \times M$ data samples for testing and save it in T_{test} dataset
   e) Using T_{train} train the SVM
   f) After training, the learning performance of SVM is evaluated. Using T_{test} , detection accuracy, false positive rate, false negative rate and time taken to run the model are measured.
4. Repeat the procedure from 2-3 until highest detection accuracy, low false positive and false negative rate are achieved. After this the randomly selected n features are stored.

At first, proposed scheme defines the number of features in the dataset. Furthermore, n features are selected using simulated annealing to generate a combination of three features among the total 47 features to see which combination of the features is relevant to achieve highest detection accuracy. After that, the algorithm selects the N number of data samples which contains both normal and abnormal data traffic pattern to run the SVM-based scheme. Then it randomly selects n number of appropriate features for detecting abnormal behavior from network data traffic. Before running the algorithm parameters of the proposed algorithm such as $gamma, coef\ \theta, nu\ etc$ were initialised. After mixing up the dataset, $n \times N$ data samples are selected so that the SVM can learn the dataset without any bias. These samples are stored in T_{train}that will be used for training purpose. Similarly $n \times M$ data samples are chosen and stored in $T_{test}$to verify the learning performance of the SVM-based detection scheme. After performing the learning procedure of SVM, detection accuracy, false positive rate and time to run the algorithm are measured. The whole process is repeated from steps 2-3 until all possible combination (kCn) of features are evaluated to analyze which combination of features provide the highest detection accuracy and lower false positive and false negative rate.

To be noted that the simulated annealing is programmed in a way that it will not generate a similar or repeated combination of features. For example, among the 47 features if a combination of feature number [1, 13, 27] is generated randomly then it will not generate a combination like [27, 13, 1] or [13, 1, 27].

5. Simulation & Results

In this section, the simulation is performed to validate the performance of the proposed algorithm. In the first phase, the experiment is conducted and analyzed that whether the proposed machine learning algorithm is able to differentiate between normal and anomaly behavior or not. The percentage of detection accuracy, false positive, false negative and time have been evaluated for the proposed method. Finally, we have investigated the performance of the proposed algorithm by randomly selecting 3 features at a time among 47 features and apply SVM on a different number of feature combinations. The whole experiment has been conducted using lib-linear machine learning tool.

From the dataset, 150,000 samples are selected randomly which contains 75,000 normal and 75,000 anomaly samples. 80% of the number of samples are used for training and rest of them are used for testing the algorithm.

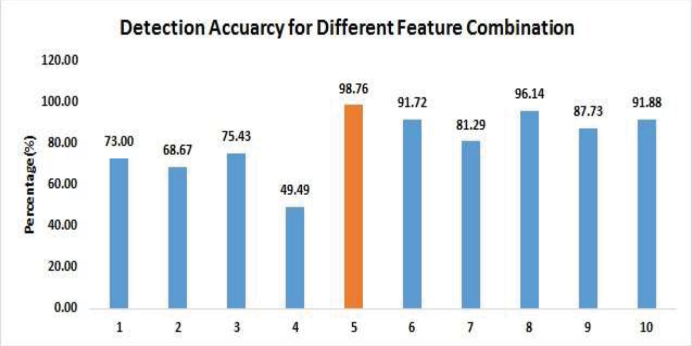

Figure 2: Percentage of Detection Accuracy

Figure 2 represents the detection accuracy of the proposed algorithm according to feature combination. In Table 1 the combination number denotes which three features were selected for that particular type combination. The proposed algorithm achieved the highest detection accuracy recorded as 98.76% when combination number 5 were selected (Please see table 1). This combination contains three important features such as time to live value, TCP advertisement value, and the number of connections that contain same service and source address in previous 100 connections The lowest detection accuracy among the given results were recorded as 49.49% when combination number 4 were selected. So feature selection works as a contributing factor in increasing the intrusion detection accuracy.

Furthermore, the performance matrix of the proposed algorithm were analyzed. The false positive refers to a situation that there is an intrusion in the system but in reality it's not an intrusion. In figure 3 the percentage of false positive and in figure 4 the percentage of a false negative is shown for the proposed scheme.

The lowest false positive and false negative were recorded as 0.09% and 1.15% respectively while combination number 5 were taken into account. So it can be inferred that while the correlative features are selected the false positive rate decreases and it represents a reliable intrusion detection system.

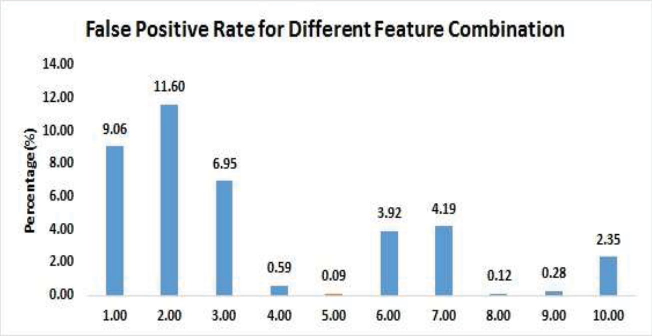

Figure 3: Percentage of False positive rate

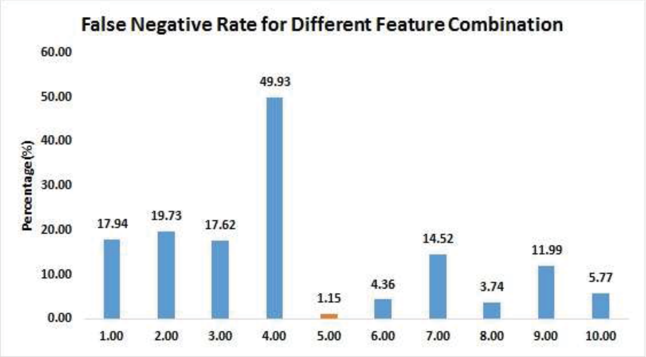

Figure 4: Percentage of False negative rate

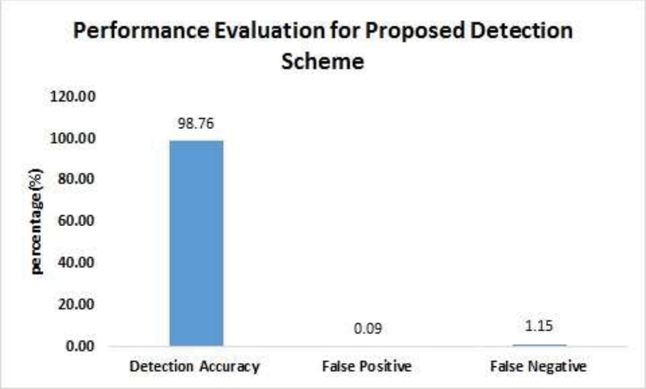

Figure 5: Performance Evaluation of the Scheme

Figure 5 represents the performance of the proposed scheme. The combination of time to live value, TCP advertisement value and the number of connections that contain same service and source address in previous 100 connections provided the highest detection accuracy with a very low false positive and false negative rate. Also using this feature combination the algorithm took only 13.84 seconds to detect an anomaly in the network traffic where the normal SVM method took 220.24 seconds to detect any anomalous behavior in the network.

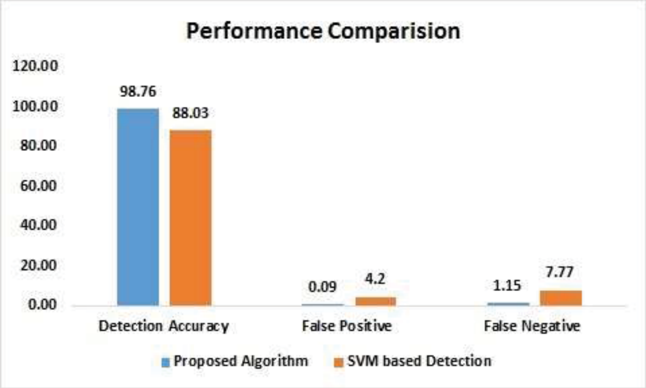

Figure 6: Performance Comparison

Figure 6 represents the comparison between the proposed algorithm and general SVM based detection scheme. As discussed before, in our algorithm we applied Simulated Annealing first to generate a combination of randomly selected 3 features and then SVM were applied. The normal SVM based scheme shows a detection accuracy of 88.03% only but it's hard to define which features needs to be taken into account to provide higher detection accuracy. Our proposed scheme provides 98.76% anomaly detection accuracy with lower false positive and false negative rate using randomly three features only.

In this research, a set of three randomly selected feature were used to evaluate the performance. In future work, a different number of randomly selected features will be taken into account to evaluate the performance of the proposed algorithm.

6. Conclusions

In this paper, a combination of two machine learning method was used for network intrusion detection. The proposed algorithm provided significant detection accuracy of 98.76% and lower false positive rate of 0.09% and false negative rate of 1.15%, whereas the normal SVM based scheme achieved a detection accuracy of 88.03% and false positive rate of 4.2% and false negative rate of 7.77%.One of the important matter is the feature selection on which the most portion of the detection accuracy depends. Further research work can be done using a combination of a very low number of features that can reduce the time to detect an anomaly in the network traffic. Furthermore, Artificial Neural network will be applied to the dataset to evaluate the performance and compare with the proposed detection algorithm.

References

[1] D. Batchelder, J. Blackbird, P. Henry, and G. MacDonald, "Microsoft Security Intelligence Report - Volume 17," *Microsoft Secur. Intell. Rep.*, vol. 16, pp. 1–19, 2014.

[2] K. Wang and S. Stolfo, "Anomalous payload-based network intrusion detection," *Recent Adv. Intrusion Detect.*, pp. 203–222, 2004.

[3] G. Xiaoqing, G. Hebin, and C. Luyi, "Network intrusion detection method based on Agent and SVM," *2010 2nd IEEE Int. Conf. Inf. Manag. Eng.*, pp. 399–402, 2010.

[4] R. P. Lippmann and R. K. Cunningham, "Improving intrusion detection performance using keyword selection and neural networks," *Comput. Networks*, vol. 34, pp. 597–603, 2000.

[5] E. Denning, R. Ave, and M. Park, "Attempted break-in --," pp. 118–131.

[6] W. Hu, Y. Liao, and V. R. Vemuri, "Robust anomaly detection using support vector machines," *Proc. Int. Conf. Mach. Learn.*, pp. 282–289, 2003.

[7] Z. Zhang, J. Li, C. N. Manikopoulos, J. Jorgenson, and J. Ucles, "HIDE: a Hierarchical Network Intrusion Detection System Using Statistical Preprocessing and Neural Network Classification," *Proc. IEEE Work. Inf. Assur. Secur.*, pp. 85–90, 2001.

[8] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," *Adv. Kernel MethodsSupport Vector Learn.*, vol. 208, pp. 1–21, 1998.

[9] Y. LeCun, L. D.Jackel, L. Bottou, A. Brunot, C. Cortes, J.~S.~Denker, H.~Drucker, I. Guyon, U. A. Müller, E. Säckinger, P. Simard, and V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," *Proc. {ICANN'95} - -International Conf. Artif. Neural Networks*, vol. II, pp. 53–60, 1995.

[10] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 130–136, 1997.

[11] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," *Comput. Vis. Pattern Recognition, 1997. Proceedings., 1997 IEEE Comput. Soc. Conf.*, pp. 193–199, 1997.

[12] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)."Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

[13] Moustafa, Nour, and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set." Information Security Journal: A Global Perspective (2016): 1-1 4.