Prediction of Network Anomaly Detection through Statistical Analysis

Abrar A. Qureshi[a,], Kamel Rekab[b]

[a]Department of Mathamatics and Computer Science, University of Virginia-Wise, Wise, VA24293, USA [b]Department of Mathematics and Statistics, University of Missouri-Kansas city, KC, MO 64110, USA
aqureshi@uvawise.edu, rekabk@umkc.edu

Abstract - Homeland security concerns continue to grow; protecting the network infrastructure remains a vital priority for government organizations as well as their private sector partners. In this paper we will focus on one-at-a-time Network Intrusion detection. Our goal is to build a Network Intrusion detection model through statistical analysis. We examined TCP/IP packet headers anomalies to predict if an intrusion is occurring or not. This approach, in turn, will provide the model that predicts the number of intrusions by maximizing the true positives ratio (real intrusions) while keeping the false positives (false alarm) ratio small. The resulting model will detect future intrusions more effectively and to protect the valuable network resources at large. The outcome of this research is validated through statistical measures such as model chi-square, its model significance (P-value), and overall model fitness. It can also be verified through ROC curves.

Keywords: Network Security, Intrusion Detection, Anomaly Detection, Logistic Regression

1. Introduction

Despite the best efforts of security professionals, networks are subjected to an increasing number of sophisticated attacks. The proliferation of cracking activity on the Internet has led to astounding developments in intrusion detection technology. Our research encompasses Prediction of Network Anomaly Detection through a novel statistical technique. Logistic regression is a technique for analyzing problems where there are one or more independent variables, which determines an outcome that is measured with a dichotomous variable in which there are only two possible outcomes. In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1(intrusion) or 0 (no intrusion).

1.1. TCP/IP Header Anomalies

TCP/IP is the world's de facto communications protocol. IP operates on gateway machines that move data from department to organization to region and then around the world. [1]. There are several TCP/IP Header Anomalies such as IP Header length is less than the minimum IP Header length (20 octets). TCP Header length is less than the minimum TCP Header length (20 octets). IP standard violations, Spoofed IP address, overlapping Offset field values as appears in Teardrop attack, SYN packet include destination IP address as a source IP address, use of reserve TCP port numbers, Invalid TCP flags for instance combination of SYN with PUSH, RST, and FIN [2].

2. Network Intrusion Detection Architectures

Intrusion detection systems generally fall into one of two categories, Anomaly-Based Detection orSignature-Based Detection.

Anomaly-based IDSs establish baselines of normal behavior by profiling particular users or network connections and then monitoring for activities that deviate from the baseline.Any network-based intrusion-detection system currently implemented, about 90% of all detects are false positives such as RealSecure, the NID, Snort, and Shadow [3].Misuse detection (signatures) model relies on comparison of traffic to a database containing signatures of known attack methods.

3. IDS Research

The proliferation of cracking activity on the Internet has led to astounding developments in intrusion detection technology. Intensive research has been done in this area. Other researchers have had similar objectives but different approaches. Paul Barford, Jeffery Kline, David Plonka and Aves Ron use wavelet filters for exposing the details of both ambient and anomalous traffic, "A Signal Analysis of Network Traffic Anomalies" [4]. Matthew Mahoney and Philp Chan used packet header anomaly detector (PHAD) detected attacks by learning the normal range of values for 33 fields of the Ethernet, IP, TCP, UDP and ICMP protocol "PHAD: Packet Packet Header Anomaly Detecting for Identifying Hostile Network Traffic" [5]. Sverre Andersen, Hallgrim Flatland, and Torbjørn Meland have used Python script to identify attacks on "Network Intrusion Detection-Instance Based learning" [6]. R. Sekar, A. Gupta, J Frullo, T. Shanbhag, A. Tiwari, H. Yang and S. Zhou combined specification-based and anomaly–based intrusion detection, and used state-machines specifications of network protocols, and augments these state machines with information about statistics that need to be maintained to detect anomalies "Specification-based Anomaly Detection: A new Approach for Detecting Network Intrusions [7].

4. Logistic Regression Approach

There are a variety of statistical techniques that can be used to predict a binary dependent variable from a set of independent variables. Multiple regression analysis and discriminate analysis are two related techniques that quickly come to mind. However, these techniques pose difficulties when the dependent variable can have only two values an event occurring (Intrusion is happening) or not occurring (Intrusion is not happening).

Outcome variable determines choice of model if the desired outcome is continuous then linear regression model is suitable. Since our desired outcome is binomial then Logistic regression is the best choice. We used Logistic regression to model the relationship between a categorical response variable and one or more predictor variables, since our data contains a mixture of numerical and categorical variables. It produces a formula that predicts the probability of the occurrence of an event as a function of the independent variables [9, 11]. The logistic model is popular because of the logistic function f (z). The model provides estimates that must lie in the range of between 0 and 1. Our aim is to develop a sound mathematical model to predict the anomaly based network intrusions with a maximum true positive and a minimal amount of false alarms.

The function, called f (z), is given as follows:

$$f\,(z) = 1/(1+e^{-Z})$$

5. Logistic Regression Statistical Analysis

In this paper we will focus on one-at-a-time Intrusion Detection. We, first check the model against a set of 14052 packets (training set) using actual test cases outcome (1 or 0). Then we validate the model against one packet (anomalous or non-anomalous) that is outside the training set and were not used to determine the model. A significant p value and a chi-square statistic are applied to test the model fitness.

Case Study

5.1. The Predictor Variables

Logistic model consists of 15-predictor variable, metric/non-metric and binary dependent variable. As an example TCP/IP Headers field variables are used in this case study. Their abbreviated names, description and anomaly model operational definitions are as follows.

Table 1

IP Header predictor variables with their values

Test parameters (Intrusion=0)	Value 1 (Intrusion=1)	Value 2
Type of service: IP_TOS		
Length: IP_LEN		
Identification: IP_ID		
Flags: IP_FLAGS		
Time to live: IP_TTL		
Header checksum: IP_CSUM		

Table 2

TCP Header predictor variables with their values

Test parameters (Intrusion=0)	Value 1 (Intrusion=1)	Value 2
Source port number: TCP_SPOR		
Destination port number: TCP_DPOR		
Sequence number: TCP_SEQ		
Acknowledgement number: TCP_ACK		
Offset: TCP_OFF		
Reserved bits: TCP_RES		
Flags (URG, ACK, PSH, RST, SYN and FIN): TCP_FLAG		
Window size: TCP_WIN		
TCP header checksum: TCP_CSUM		

5.2. The Logistical Regression Model Analysis

This analysis involves estimation of a logistical regression model using block entry of variables. This study seeks to predict intrusions based on the TCP/IP packet headers anomalies. Dataset used for this research was captured from Targa when the system was not under attack therefore there are few intrusions in this particular dataset. The total number of TCP/IP packets involved in this research was 16000 in which 1533 were found abnormal (intrusions=1), and 14467 packets were normal (intrusions=0). IP version 4.0 was used in all the packets detected. In this paper, we detected one-at-a-time intrusion and propose a test for anomaly detection based on standard statistical tests. These include the logistic Regression and Receiver Operation Characteristic Curve. We were able to predict anomalies while minimizing false alarms and maximizing intrusion detection.

5.3. Prediction of single anomalous (with intrusion) packet

We created the best model that predicts intrusions based on a set of 14052 packets. The model's performance was then measured by comparing the predicted intrusions and the observed intrusions (Tables 4-6). Cross validation was based on 1 packet that was not used to determine the model. Selected packet was among the 1533 packets with intrusions, and we perfectly predicted that packet, with an undetermined false alarm (Table 7).

5.4. Prediction of single non-anomalous (without intrusion) packet

The same model was used to predict intrusions based on a set of 14052 packets. This time we selected a packet among the non-intruded packets. The model's performance was then measured by comparing the predicted intrusions and the observed intrusions (Tables 9-11). Cross validation was based on 1 packet that was not used to determine the model. Selected packet was among the 14467 packets without intrusions, and we perfectly predicted that packet, with an undetermined false alarm (Table 12).

5.5. Relation Between the Most Significant TCP/IP Protocol Characteristics and the Probability of Intrusion

The logistic regression model was performed to predict the probability that an intrusion occurs given a set of Internet protocol characteristics. The model Chi-square is very significant (P value =. 0000), Table 4 and 9 (sig. 0). It shows that our predictive model performs very well.

The logistic model given by the prediction equation was used to predict one-at-a-time intrusions on a new set of data that consists of 14052 packets. The 14052 packets had 1533 number of intrusions. Our model was able to predict every intrusion one-at-a-time. Our model was also able to predict every non-intrusion one-at-a-time, which shows that our model performs very well in both cases.

The Receiver operation characteristic curve also shows that our predictive model is nearly perfect. The probability of perfection is 100 percent.

Table 3

Predictive Model Based on 14052 packets

Variables		B	S.E.
Step 1(a)	IP_TOS	4.113	2019.471
	IP_LEN	.000	.595
	IP_ID	.000	.016
	IP_FLAGS	-1.863	298.801
	IP_TTL	.011	3.901
	IP_CSUM	.000	.017
	TCP_SPOR	.000	.037
	TCP_DPOR	.000	.018
	TCP_SEQ	.000	.000
	TCP_ACK	.000	.000
	TCP_OFF	-11.858	552.223
	TCP_RES	.094	26.905
	TCP_FLAG		
	TCP_FLAG(1)	13.727	22304.241
	TCP_FLAG(2)	14.244	22295.835
	TCP_FLAG(3)	31.789	22610.362
	TCP_FLAG(4)	40.902	22380.180
	TCP_FLAG(5)	-3.159	28396.043
	TCP_FLAG(6)	13.497	22525.803
	TCP_FLAG(7)	11.602	23736.032
	TCP_FLAG(8)	13.144	24306.246
	TCP_FLAG(9)	8.989	24558.579
	TCP_WIN	.000	.018
	Constant	24.016	22137.452

Table 4

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	235.258	23	.000
	Block	235.258	23	.000
	Model	235.258	23	.000

Table 5

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000	.017	1.000

Table 6

Selected Cases

	Observed		Predicted		
			Selected Cases		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	14037	0	100.0
		1	0	15	100.0
	Overall Percentage				100.0

Table 7 (Removed table 8)

Cross Validation

	Observed		Predicted		
			Unselected Cases		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	0	0	
		1	0	1	100.0
	Overall Percentage				100.0

Table 8

Predictive Model Based on 14052 packets

Variables		B	S.E.
Step 1	IP_TOS	4.113	2019.470
	IP_LEN	.000	.595
	IP_ID	.000	.016
	IP_FLAGS	-1.863	298.811
	IP_TTL	.011	3.902
	IP_CSUM	.000	.017
	TCP_SPOR	.000	.037
	TCP_DPOR	.000	.018
	TCP_SEQ	.000	.000
	TCP_ACK	.000	.000
	TCP_OFF	-11.858	552.237
	TCP_RES	.094	26.913
	TCP_FLAG		
	TCP_FLAG(1)	13.687	19627.177
	TCP_FLAG(2)	14.204	19616.679
	TCP_FLAG(3)	31.749	19960.050
	TCP_FLAG(4)	40.862	19716.540
	TCP_FLAG(5)	-3.199	26335.118
	TCP_FLAG(6)	13.457	19875.932
	TCP_FLAG(7)	11.563	21241.530
	TCP_FLAG(8)	13.105	21875.820
	TCP_FLAG(9)	8.950	22154.900
	TCP_WIN	.000	.018
	TCP_CSUM	.000	.017
	Constant	24.056	19423.935

Table 9

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	248.876	23	.000
	Block	248.876	23	.000
	Model	248.876	23	.000

Table 10

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000	.018	1.000

Table 11

Selected Cases

	Observed		Predicted		
			Selected Cases		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	14036	0	100.0
		1	0	16	100.0
	Overall Percentage				100.0

Table 12

Cross Validation

	Observed		Predicted		
			Unselected Cases		
			Y		Percentage Correct
			0	1	
Step 1	Y	0	1	0	100.0
		1	0	0	
	Overall Percentage				100.0

6. Receiver Operating Characteristic (ROC) Curve

The ROC curve plots the false positive and 1- the false negative rate on X-axis and on the Y-axis respectively. ROC curves are very useful for evaluating the predictive accuracy of a chosen model in logistic regression. The predicted values generated by the logistic model can be viewed as a continuous indicator to be compared to the observed binary response variable [8, 10].

We obtained this curve by plotting the sensitivity against 1-specificity. Smaller values (0.5 or less) indicate a weaker result whereas larger values of the test result variable(s) indicate stronger evidence for a positive actual state. Our result shows area under the ROC curve is close to 1, which indicates a very good result.

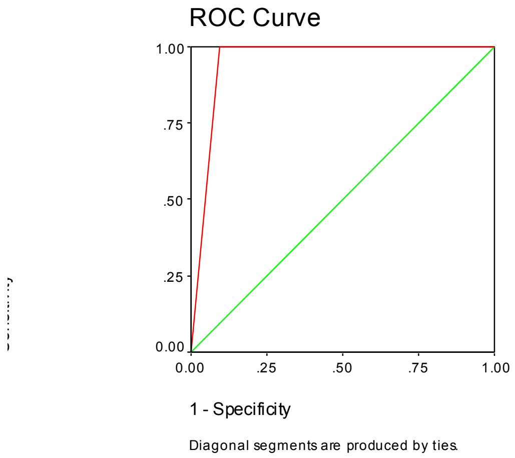

Fig. 5. ROC Curve

Area Under the Curve

Area Under the Curve
.952

7. Conclusion

Network security is no longer an option. An Intrusion Detection System is a tool that is part of a good security architecture and Multi-Layered Defense Strategy such as Security policy, Firewalls, Router security, Host system security, Auditing, Intrusion detection systems, Monitor and Response mechanism however, once the IDS is deployed onto the network the system must be monitored.

Using multiple layers in a security model is the most effective method of deterring unauthorized use of computer systems and network services. Every layer provides some protection from intrusions; thus, defeating one of the layers will not lead to compromising the security of the entire organization.

We developed a new approach for predicting Anomaly based Network Intrusion Detection through Logistic Regression. We predicted intrusions based on the TCP/IP packet headers anomalies and were able to predict anomalies while minimizing false alarms and maximizing intrusion detection. We created the best model that predicts one-at-a-time intrusion based on a set of 14052 packets. The model's performance was then measured by comparing the predicted intrusions and the observed intrusions. We perfectly predicted all the intrusions, without false alarm. Predictive accuracy of a chosen model is evaluated by the ROC curve. The area under the curve is 0.952, which is close to 1 and represents the perfect indicator. We will examine large data sets with TCP/UDP/IP from different sources in our future research to validate our results further.

References

[1] W. Richard Stevens. TCP/IP Illustrated Volume 1, Feb 2005.

[2] Ankit Fadia,Network security - A Hacker's perspective, 2003.

[3] Stephen Northcutt and Judy Novak, Network Intrusion Detection, 2001.

[4] Paul Barford, Jeffery Kline, David Plonka and Aves Ron, "A Signal Analysis of Network Traffic Anomalies", ACM 2002 ISBN 1-58113-603-x/02/0011.

[5] Matthew Mahoney and Philp Chan "PHAD: Packet Packet Header Anomaly
Detecting for Identifying Hostile Network Traffic" Florida Tech Technical Report CS-2001-04.

[6] Sverre Andersen, Hallgrim Flatland, and Torbjørn Meland, Network Intrusion Detection –Instanced Based Learning, IKT2340 - Open Systems, seminar (2003).

[7] R. Sekar, A. Gupta, J Frullo, T. Shanbhag, A. Tiwari,H. Yang and S. Zhou, Specification-based Anomaly Detection: A new Approach for Detecting
Network Intrusions, ACM 2002 ISBN 1-58113-612-9/02/0011.

[8] D.G. Kleinbaum, L.L. Kupper, Applied Regression Analysis and Other Multivariable Methods, Duxbury Press, North Scituate, New York, 1976.

[9] B.W. Lindgren, Statistical Theory, third ed., Macmillan, New York, 1976.

[10] Kamel Rekab, Muzaffar Shaikh, Statistical Design Of Experiments With Engineering Applications - CRC Press (April 8, 2005).

[11] Hosmer, D. W. and Lemeshow, S. Applied Logistic Regression, John Wiley, 1989.