

# Anomaly Detection with Multinomial Logistic Regression and Naïve Bayesian

Nguyen Dai Hai and Nguyen Linh Giang

**Abstract** Intrusion Detection by automated means is gaining widespread interest due to the serious impact of Intrusions on computer system or network. Several techniques have been introduced in an effort to minimize up to some extent the risk associated with Intrusion attack. In this paper, we have used two novel Machine Learning techniques including Multinomial Logistic Regression and Naïve Bayesian in building Anomaly-based Intrusion Detection System (IDS). Also, we create our own dataset based on four attack scenarios including TCP flood, ICMP flood, UDP flood and Scan port. Then, we will test the system's ability of detecting anomaly-based intrusion activities using these two methods. Furthermore we will make the comparison of classification performance between the Multinomial Logistic Regression and Naïve Bayesian.

**Keywords** DoS · Logistic regression · Naïve Bayesian · Intrusion detection system

## 1 Introduction

Intrusion Detection is a process of gathering intrusion related knowledge that occurred in the computer networks or systems and analyzing them for detecting future intrusions. Intrusion Detection can be divided into two categories: Anomaly

---

N. D. Hai (✉)

School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam  
e-mail: haidnguyen0909@gmail.com

N. L. Giang

Department of Communication and Computer Networks, Hanoi University of Science and Technology, Hanoi, Vietnam  
e-mail: giangnl@soict.hut.edu.vn

detection [2] and Misuse detection. The former analyses the information gathered and compares it to a defined baseline of what is seen as “normal” service behaviors, so it has ability to learn how to detect network attacks that are currently unknown. Misuse detection is based on signatures for known attacks, so it is only as good as the database of attack signatures that it uses for comparison. Misuse detection has low false positive rate, but can not detect novel attacks. However, anomaly detection can detect unknown attacks, but has high false positive rate.

The Naïve Bayesian (NB) method is based on the work of Thomas Bayesian. 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.

A Multinomial Logistic Regression (MLR) model is used for data in which the dependent variable is unordered or polytomous, and independent variables are continuous or categorical predictors. This type of model is therefore measured on a nominal scale and was introduced by McFadden (1974). Unlike a binary logistic model in which a dependent variable has only a binary choice (e.g., presence/absence of a characteristic), the dependent variable in a multinomial logistic model can have more than two choices that are coded categorically, and one of the categories is taken as the reference category.

In this paper, we propose two methods MLR and NB in building anomaly-based IDS and compare the performance of two linear classifier of Naïve Bayesian (NB) and multinomial Logistic Regression (MLR) based on attack scenarios which we created, and search for the characteristics of the data that determine the performance. The comparison between LR and MNB has been studied theoretically by Ng and Jordan (2002).

This paper is organized as follows: [Sect. 2](#) deals with the description of data set for our experiment. [Section 3](#) deals with foundation of methods including naïve Bayesian, multinomial logistic regression, In this section we will consider the problem of applying the two methods in building anomaly-based IDS. In [Sect. 4](#), we give an illustration and experimental results with four attack scenarios. It help in understanding of this procedure, a demonstrative case is given to show the key stages involving the use of the introduced concepts. [Section 5](#) is conclusion.

## 2 Dataset

Our data set is created by the following activities:

**Data collection activity:** collection attribute-value of the flow in terms of packet data (IP, port, TCP, UDP, ICMP). Based on these attributes, the program will build Profile (bin level) which contains the characteristic parameters for network traffic in a given time, including: (1–2) Entropy compression rate of the source/destination IP address, (3–4) Entropy compression rate of the source/destination port, (5) number of packets, (6) total size of the packets, (7) average size of packets, (8) standard

deviation of packet size, (9) number of TCP packets, (10) number of UDP packets and (11) number of ICMP packets.

**Statistical analysis activity:** This activity is based on the data have been analyzed from the data collected to build the corresponding bin arrays. The bin is divided into the following levels: hours, days, months correspond to the three classes of data is the current class, reference class and the differential classes:

**Cur\_bin:** represent for each instance “bin” (bin is the smallest time unit, in my program one minute). These instances is continuously created in the processes monitoring network traffic.

**Ref\_bin:** represents the reference model corresponding to one unit of time reference. Reference model is adaptably updated, based on values of Cur\_bin in the absence of intrusion detection.

**Dif\_bin:** represents the difference between the current value and the reference value and is the input of classifiers.

## 3 Methods

### 3.1 Naïve Bayesian

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. Naïve Bayesian classifiers allow the representation of dependencies among subsets of attribute [9]. Through the use of Bayesian networks has proved to be effective in certain situations, the result obtained, are highly dependent on the assumption about the behavior of the target system, and so a deviation in these hypotheses leads to detection errors, attributable to the model considered [10]. The NB classifier work as follows: Let T be a training set of samples, each with their class labels. There are k classes  $C_1, C_2, \dots, C_k$ , each sample is represented by an n-dimensional vector  $X = \{X_1, X_2, \dots, X_n\}$ .

Given a sample X, The classifier will predict that X belongs to the class having the highest a posteriori probability, conditional on X. That is X is predicted to belong to the class C, if and only if  $P(C_i|X) > P(C_j|X)$  for  $1 \leq j \leq m, j \neq i$ .

By bayes' theorem, we have  $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$ . As P(X) is the same for all classes and only  $P(C_i)$  are not known, then it is commonly assumed that the classes are equally likely, that is,  $P(C_1) = P(C_2) = \dots = P(C_m)$  we would therefore maximize  $P(X|C_i)$ .

In order to reduce computation in evaluating  $P(X|C_i)$ . The naïve assumption of class conditional independence is made. Mathematically this means that  $P(X|C_i) \approx \prod_{k=1}^n P(X_k|C_i)$ . The probabilities  $P(X_k|C_i)$  can easily be estimated from the training set. If X is continuous-valued, then we typically assume that the values have a Gaussian distribution with a mean  $\mu$  and standard deviation  $\sigma$ . So that

$P(X_k|C_i) = g(X_k, \mu_{ci}, \sigma_{ci})$ . We need to compute  $\mu_{ci}, \sigma_{ci}$  in training stage. In order to predict the class label of  $X$ ,  $P(X|C_i)P(C_i)$  is evaluated for each class  $C_i$ . The classifier predicts that the class label of  $X$  is  $C_i$  if and only if it is the class that maximizes  $P(X|C_i)P(C_i)$ .

### 3.2 Multinomial Logistic Regression

A multinomial logistic regression model is used for data in which the dependent variable is unordered or polytomous, and independent variables are continuous or categorical predictors. This type of model is therefore measured on a nominal scale and was introduced by McFadden (1974). Unlike a binary logistic model in which a dependent variable has only a binary choice (e.g., presence/absence of a characteristic), the dependent variable in a multinomial logistic model can have more than two choices that are coded categorically, and one of the categories is taken as the reference category. This study used “0” (normal) as the reference category. Suppose  $y_i$  is the dependent variable with five categories for individual connection  $i$ -th, and the probability of being in category  $s$  ( $s = “1”$  [TCP flood], “2” [ICMP flood], “3” [UDP flood], “4” [Scan Port]) can be denoted  $\pi_i^{(s)} = \Pr(y_i = s)$  with the chosen reference category,  $\pi_i^{(0)}$ . Then, for a simple model with one independent variable  $x_i$ , a multinomial logistic regression model with logit link can be represented as:

$$\log\left(\frac{\pi_i^{(s)}}{\pi_i^{(0)}}\right) = \beta_0^{(s)} + \beta_1^{(s)}x_i, s = 1, 2, 3, 4.$$

An alternative way to interpret the effect of an independent variable,  $x$ , is to use predicted probabilities  $\pi_i^{(s)}$  for different of  $x$ :

$$\pi_i^{(s)} = \frac{\exp(\beta_0^{(s)} + \beta_1^{(s)}x_i)}{1 + \sum_{k=1}^4 \exp(\beta_0^{(k)} + \beta_1^{(k)}x_i)}.$$

Then, the probability of being in the reference category, “0” (normal), can be calculated by subtraction:

$$\pi_i^{(0)} = 1 - \sum_{k=1}^4 \pi_i^{(k)}$$

## 4 Experiment and Results

In this section, we summarize our experimental results to detect network intrusion detections using Naïve Bayes and Multinomial Logistic Regression over dataset we created based on four attack scenarios including: TCP flood, ICMP flood, UDP flood and Port Scan.

### 4.1 Purpose of Study

The objective of this study is to detect some common attack types in computer systems and networks. We furthermore make the comparison of classification performance between the NB and MLR model.

### 4.2 Dataset

In this study, the measured attributes are (in particular, 11 attributes): entropy compression rate of the source/destination IP address and source/destination port, number of packets, total/average size of the packets, standard deviation of packet size and number of TCP/UDP/ICMP packets, So each instance will be represented by a vector including 11 attributes and the input of each classifier is differential vector of current vector and reference vector which refer to normal state (Table 1).

### 4.3 Experiment

We will test the system's ability of detecting anomaly-based intrusion activities using two methods: Naïve Bayes and Multinomial Logistic Regression. We will proceed on the four attack scenarios including ICMP flood, TCP flood, UDP flood and port scan. Using with each attack will change significantly the number of ICMP, TCP, UDP packets and entropy source/target.

**Table 1** Number of examples in dataset we created

Attack types	Training samples
Normal	110
TCP flood	205
ICMP flood	200
UDP flood	150
Scan port	180

### 4.3.1 Testing Environment

The system was tested on virtual LAN 100 Mps environment using VMware tool, including two Window XP computers and a Ubuntu computer installed the Anomaly IDS. These computers are connected to each other through a virtual switch.

### 4.3.2 Testing Scenarios

Two Window XP computers implement TCP flood, UDP flood, ICMP flood refer to bandwidth flood attacks using tools like hping3, udpflood.exe, ping respectively or scan port in range 1–300 on Ubuntu computer installed anomaly IDS. Our program will collect and analysis packets in order to detect anomalous in traffic.

### 4.3.3 Experimental Results

A “confusion matrix” is sometime used to represent the result of, as shown in Table 2 (Naïve Bayes) and Table 3 (Multinomial Logistic Regression). The advantage of using this matrix is that is not only tells us how many got misclassified but also what misclassification occurred. We define the Accuracy, Detection rate and false-alarm:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad Detection - rate = \frac{TP}{TP + FP}$$

$$False - Alarm = \frac{FP}{FP + TN}$$

FN: False Negative, TN: True Negative, TP: True Positive and FP: False Positive (Table 4).

**Table 2** Confusion matrix for naïve bayes

Actual	Predicted normal	Predicted TCP flood	Predicted ICMP flood	Predicted UDP flood	Predicted scan port	Accuracy (%)
Normal	110	0	0	0	0	100
TCP flood	1	201	0	0	3	98
ICMP flood	2	0	197	1	0	98.5
UDP flood	0	0	1	147	2	98
Scan port	1	2	1	4	172	95.6

**Table 3** Confusion matrix for multinomial logistic regression

Actual	Predicted normal	Predicted TCP flood	Predicted ICMP flood	Predicted UDP flood	Predicted scan port	Accuracy (%)
Normal	110	0	0	0	0	100
TCP flood	0	204	0	0	1	99.5
ICMP flood	1	0	203	1	0	99
UDP flood	0	0	0	149	1	99.3
Scan port	1	0	0	1	177	98.33

**Table 4** Comparison between BN and MLR

	Naïve bayes		Multinomial logistic regression	
	Detection rate	False alarm	Detection rate	False alarm
Normal	100	1	100	1
TCP flood	99	0.33	100	0
ICMP flood	98.99	0.4	100	0
UDP flood	96.7	0.625	98.7	0.66
Port scan	97	0.45	98.8	0.5

## 5 Conclusion

This study constructed an Anomaly-based Intrusion Detection Model based on Naïve Bayes and Multinomial Logistic Regression algorithm. We also experiment IDS's ability of detection using both these methods in the data sets that we created based on four attack scenarios including ICMP flood, UDP flood, TCP flood and Scan Port. The experimental results show that both two methods give very high accuracy and could be applied in practice. However, this is still only the initial test, and more research is needed, in the future we will continue to improve and experiment in a real network environment.

## References

1. Lippmann R, Haines JW, Fried DJ, Korba J, Das K (2000) The 1999 DARPA off-line intrusion detection evaluation. *Comput Netw* 34:597–595
2. Stillerman M, Marceau C, Stillman M (1999) Intrusion detection for distributed systems. *Commun ACM* 42(7):62–69
3. Chang CC, Lin CJ (2009) LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. 18th November 2009
4. Anderson J (1980) Computer security threat monitoring and surveillance. James P. Anderson Co, Washington
5. Yu Y, Hao H (2007) An ensemble approach to intrusion detection based on improved multi-objective genetic algorithm. *J Softw* 18(6):1369–1378

6. Luo J, Bridges SM (2000) Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. *Int J Intell Syst* 15(8):687–703
7. Barbard D, Wu N, Jajodia S (2001) Detecting novel network intrusions using bayes estimator. In: *Proceeding of the 1st SIAM international conference on data mining*
8. Kuchimanchi G, Phoha V, Balagani K, Gaddam S (2004) Dimension reduction using feature extraction methods for real-time misuse detection systems. In: *Fifth annual IEEE proceedings of information assurance workshop*, pp 195–202
9. Han J, Kamber M, (2012) *Data mining: concepts and techniques*. Elsevier, San Francisco
10. Garcia-Teodoro P, Díaz-Verdejo JE, Maciá-Fernández G, Vázquez E (2009) Anomaly-based network intrusion detection: techniques, systems and challenges. *Comput Secur* 28(1–2):18–28
11. Phoha VV (2002) *The springer dictionary of internet security*. Springer, New York
12. Vapnik VN (1999) *Statistical learning theory*. Wiley-Interscience, New York