Fjrs.futa.edu.ng
FUTA Journal of Research in Sciences
ISSN: 2315 – 8239 (Print); E-ISSN: 2489 - 0413

FUTA Journal of Research in Sciences, Vol. 13 (2), October, 2017: 364- 370

BAYESIAN APPROACH to NETWORK INTRUSION DETECTION on WEB SERVER LOG DATA

[1]B. A. Onyekwelu, **[2]A. O. Adetunmbi**

[1]Department of Mathematics and Computer Science, Elizade University, Ilara-Mokin, Ondo State, Nigeria.
[2]Department of Computer Science , Federal University of Technology, Akure, Ondo State, Nigeria
Corresponding author's email: bukola.onyekwelu@elizadeuniversity.edu.ng

ABSTRACT
The astronomical growth in global interconnectivity is a highly welcomed development, due to its positive impact on schools, organizations, institutions and individuals, especially with respect to access to critical information. However, this growth has also brought about increase in unauthorized access to critical enterprise information, leading to unforeseen risks in information management. This paper first discussed the Bayesian approach to Intrusion Detection, and the Architectural Model of a Bayesian Rate Intrusion Detection System. Preprocessed data obtained from Web Server Log of a University Network was used. Label Extraction and Data Set Balancing were carried out on the preprocessed data. Training and testing datasets where obtained, and a Bayes classifier, based on the Bayesian Information Criterion, was implemented on a System running Windows Operating system, using C++ programming language. The class labels were compared from the results obtained. it is clear from the results that Bayesian approach takes care of the problem of false alarms, and therefore, it is a standard tool for Intrusion Detection. The tools provide a means of solving the problems of data and information security.
Keywords: Intrusion Detection, Discretization, Bayesian Information Criterion,

INTRODUCTION
Global Interconnectivity is a phenomenon that is presently growing astronomically, as more schools, organizations, private companies, governmental institutions and even individuals are connecting for critical information search, processing and dissemination. This interconnectivity allows for better productivity, faster communication capabilities and immeasurable personal conveniences. However, increased interconnectivity has its own share of limitations. It opens the door to many unforeseeable risks in Information management due to the fact that fraudulent individuals are increasingly gaining unauthorized access to critical enterprise information. Available means of preventing and detecting network infringements are becoming insufficient at a very fast rate. This is evident in the fact that recent years have witnessed a phenomenal increase in network attack incidents. This calls for regular research into the techniques of Intrusion detection and prevention

REVIEW OF RELATED WORKS
Vijayarani and Sylviaa (2015) described Intrusion Detection System as an application that is used to monitor the network and protect it from intruders. They likened an Intrusion Detection System to a burglar alarm system. Intrusion Detection systems are similar to Intrusion Prevention Systems, but differ in their operations, as described in Ashoor, and Gore (2011). They explained that while IDS is used majorly for detecting threats or intrusions, IPS is dedicated on identifying the threats or intrusions and blocking them. The similarities however include signature matching, packet inspection, and protocol validation. Peng and Zuo (2006) employed data mining approach to network intrusion detection framework in real-time. Their framework was a distributed architecture, which comprised of sensor, data preprocessor, features

extractors and detectors. Li (2012), in his paper, defined a technique of applying Genetic Algorithm (GA) to network Intrusion Detection Systems. In this technique, both temporal and spatial information of network connections are taken into consideration, in encoding the network connection information into rules in IDS. This was found to be useful in the identification of complex anomalous behaviors. The specific area of focus for this work is the TCP/IP network protocols.

BAYESIAN APPROACH TO INTRUSION DETECTION

Adetunmbi(2008) defined the Bayesian approaches as being powerful tools used in making decision and reasoning in uncertain circumstances. They also stipulated that probabilistic concept representations are used, and range from the Naïve Bayes (NB) to the Bayesian network. According to Adetunmbi(2008), the Bayesian network is a graphical model used for probabilistic relationships among a set of variables. There are two components that define a Bayesian Network, namely, a directed acyclic graph (DAG) and a set of conditional probabilistic tables.

Van Trees(2001), as cited by Axelsson and Sands(2006), employed the Bayes criterion as a useful measurement of success in the determination of the decision rule. According to Neath and Cavanaugh(2012), the Bayesian Information Criterion was first presented by Schwarz, G. in 1978, to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. The BIC is employed in comparing non-nested models, as well as models that are based on different probabilistic distributions.

Applying the Bayesian model to the Intrusion Detection System, we consider some relative entities, as discussed below.

a. The Source: The source of intrusion, in this situation, is a human computer user who issues commands to the computer via several input devices. In reality, there are different types of intrusions, needing different detectors. This makes the problem a multi-level problems, and there is need to differentiate between H_0 and H_1, H_2, H_3,…,Hn, where H_1, H_2,…'Hn are different types of intrusions. However, in the Honeypot environment, it is expected that the source is mostly malicious; therefore, expected signal is *H1*.
b. The Probabilistic Transition Mechanism: Observation is key to detecting intrusive behavior, and various means of observation include security logging mechanism, observing the network traffic coming from the user, and monitoring user's keystrokes. Noises of varying magnitude are added to the signal by the PTM, and these noises are in turn modeled and integrated into the overall model of the system. An example of such noise is behaviors from other users.
c. Observation Space: This is made up by the set of possible observations, given a particular source and channel model. There is need to discover a coordinate transformation that will transform every information into one coordinate within the Observation Space.
d. The Decision Rule: The threshold distinguishing between *H0* and *H1* needed to be defined, probably using anomaly detection method.

A Classical Detection Model, according to Axelsson and Sands (2006), is illustrated in Figure 1.

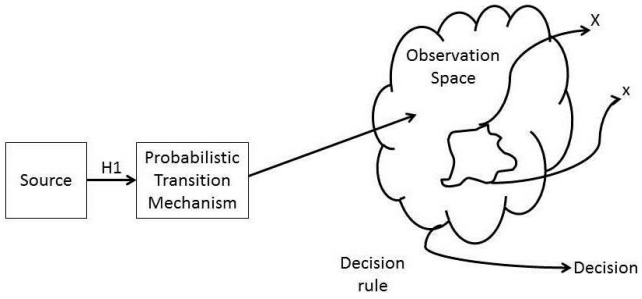

Figure 1: A Classical Detection Model [Source: Axelsson, S. and Sands, D. (2006)]

THE BASE-RATE FALLACY

The base-rate fallacy is derived from the Bayes's theorem which expresses the relationship between a conditional probability and its opposite, that is, when the condition is reversed. This is shown in the equation below

$$P(A|B) = \frac{P(A).P(B|A)}{P(B)}$$

When *P(B)* is expanded for the set of all *n* possible, mutually exclusive outcomes *A*, we get,

$$P(B) = \sum_{i=1}^{n} P(A_i) \,.\, P(B|A_i)$$

Putting the two equations together, we get,

$$P(A|B) = \frac{P(A) \,.\, P(B|A)}{\sum_{i=1}^{n} P(A_i) \,.\, P(B|A_i)}$$

Now, to calculate the Bayesian rates, we start by defining some terms:

I = intrusive behavior

$\neg I$ = non-intrusive behavior

A = presence of an intrusion alarm

$\neg A$ = absence of an intrusion alarm

The probability $P(A|I)$ – This is the probability of having an intrusion alarm when the behavior is intrusive, also known as the true positive rate, obtained during the testing of the detector against a set of scenarios that represent intrusive behavior.

The probability $P(A|\neg I)$ – the false positive rate (probability of an intrusion alarm on a non-intrusive behavior), gotten in an analogous way.

$P(\neg A|I)$ – the false negative rate, and $P(\neg A|\neg I)$ – the true negative rate can be denoted as

$$P(\neg A|I) = 1 - P(A|I)$$

$$P(\neg A|\neg I) = 1 - P(A|\neg I)$$

The focus, however, is to ensure that $P(A|I)$ and $P(\neg A|\neg I)$ remain as large as possible, that is, the probability that an alarm really indicates an intrusion, and also that the absence of an alarm signifies that there is nothing to worry about. To obtain $P(A|I)$, we apply the Bayes's theorem;

$$P(I|A) = \frac{P(I).P(A|I)}{P(I).P(A|I) + P(\neg I).P(A|\neg I)}$$

Also, for $P(\neg A|\neg I)$, we have

$$P(\neg I|\neg A) = \frac{P(\neg I).P(\neg A|\neg I)}{P(\neg I).P(\neg A|\neg I) + P(I).P(\neg A|I)}$$

As outlined in Fig 1, the Observation Space is made up by the set of possible observations, given a particular source and channel model.

THE ARCHITECTURAL MODEL

The Observation Space has been defined as the set of possible observations, given a particular source and channel model. Source here is the internet traffic, where real life data is collected. After collection, the following operations are carried out on the data;

i. Discretization – to reduce the data to manageable size
ii. Labelling – to label the data behavior as intrusive or non-intrusive
iii. Filtering – to eliminate incomplete data

The resulting data is then passed through the Bayesian Rates Equation in order to arrive at a decision. This process is illustrated in figure 2.

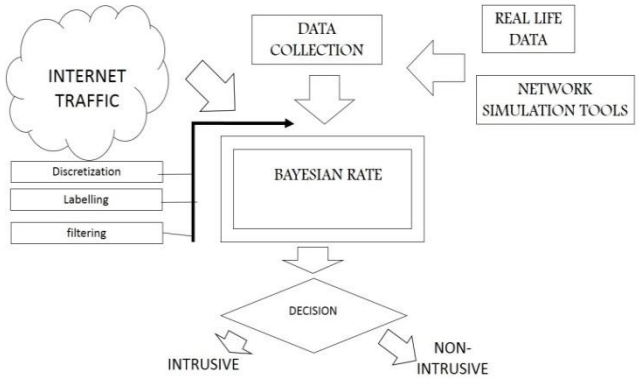

Figure 2: The Architecture Model of a Bayesian Rate Intrusion Detection System

THE NETWORK ENVIRONMENT

The Network environment under observation in this research is that of a University environment. The network is a Metropolitan Area Network (MAN), with token ring topology. Here, VSAT and Fibre Optics link via radio is used for the Internet. The University runs both wired and wireless connectivity, with several antennas, radios, routers, access points and switches spread all over the campus. The University server runs Apache 2.2.26, and is set to log according to the

general log format. The text file generated consisted of several lines, with each line showing a single HTTP access request. The request field is made up of request methods ('GET', "POST', etc), the path to the requested resource, as well as the method of access (e.g., 'HTTP 1.0). Over 1 million records were obtained, and a log reduction scheme was adopted to remove duplicates and select actual request fields.

DATA COLLECTION

The web server log obtained for this research is the Access Log for a period of three months. A large volume of web usage data was obtained from the resulting log files. This is illustrated in Figure 3.

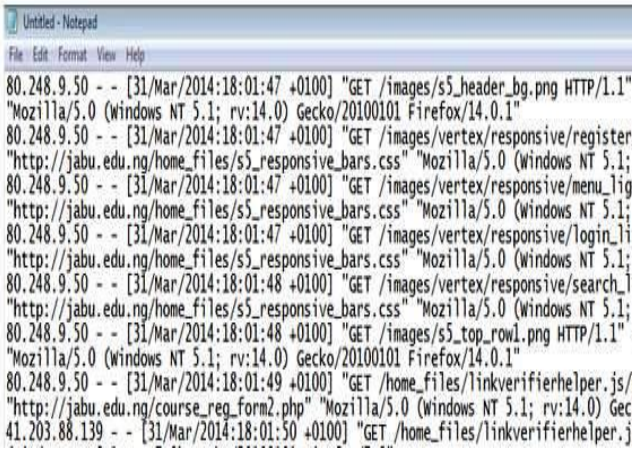

Figure 3: Text File of Log generated from Web Server

DATA PREPROCESSING

Data preprocessing involves several tasks. They are:

i. Data cleaning – where missing values are filled in, noisy data is smoothened, outliers are identified and removed, and inconsistencies are resolved.
ii. Data integration – where multiple databases, data cubes or files are integrated
iii. Data transformation – where Normalization and aggregation are carried out.
iv. Data reduction – where reduced representation is obtained in volume but the same or similar analytical results are produced.
v. Data discretization – still a part of data reduction but where particular importance is placed on numerical data.

According to Adetunmbi (2008), in a situation where data is made up of attribute values that are both or either discreet and/or continuous, there is need for discretization, before training can be carried out. In this research, data discretization becomes imperative because of the nature and volume of the data. Since the data involved in this research is web server data, the data mining carried out here can be referred to as Web Usage Mining. Grace et al (2011) explains with illustration, preprocessing as one major step involved in Web Usage Mining. The Data Preprocessing of this particular sets of data, and the results have been discussed in Onyekwelu, et al (2017).

DATA ANALYSIS

Label Extraction

From the discretized data (Onyekwelu, et al, 2017), the testing datasets are extracted for each class, and for each month. The extracted classes are compared with the predicted classes to determine the percentage accuracy.

Data Set Balancing

From the data analysis carried out in Onyekwelu, et al(2017), it is clear that there is the problem of imbalanced dataset. Imbalanced dataset affects, to a large extent, the process of machine learning. This problem, according Adetunmbi(2008), arises in a situation where we have many more occurrences of some particular classes than other classes. This affects greatly the class performance, as well as the learning outcome. In other words, a data set displaying an unequal distribution among classes can be said to be imbalanced (Ramyachitra and .Manikandan, 2014). Ramyachitra and .Manikandan(2014) proposed several approaches in handling the problem of imbalanced data set, one of which is data level approach, using the method of adaptive sampling and synthetic data generation. This method was adopted in this research. Redundant data were first removed, and synthetic data were created in the minority classes by selecting some of the nearest minority neighbors of a minority data and generating

synthetic minority data along with the lines between the minority data and the nearest minority neighbors.

RESULTS EVALUATION USING THE BAYESIAN INFORMATION CRITERION

Four distinct classes were obtained from the balanced data set; they are Normal, CSS, SQLI, and Unicode. The training data was the discretized data before balancing, while the balanced dataset was used as the testing data.

The Bayes Training and Testing program, written in C++ Programming Language, was used to evaluate the training and testing data.. This program is based on the Bayes Information Criterion is based on the BIC formular earlier stated, as follows,

$$P(I|A) = \frac{P(I).P(A|I)}{P(I).P(A|I)+P(\neg I).P(A|\neg I)} \quad (1)$$

and

$$P(\neg I|\neg A) = \frac{P(\neg I).P(\neg A|\neg I)}{P(\neg I).P(\neg A|\neg I) + P(I).P(\neg A|I)} \quad (2)$$

Where $P(I|A)$ and $P(\neg I|\neg A)$ are the true positive and true negative rates respectively. The results were classified into the four distinct classes, Normal, CSS, SQLI, and Unicode. From the balanced datasets, the Class Occurrence for the three months is shown below.

Table 1: Class Occurrences

Month 1		Month 2		Month 3	
Class	Occurrences	Class	Occurrences	Class	Occurrences
1	747	1	771	1	923
2	130	2	381	2	300
3	124	3	0	3	78
4	120	4	0	4	84
Total	1121	Total	1152		1385

Results obtained from the BIC Program

Class Probability for Month 1

$$\Pr(1) = \frac{747}{1121} = 0.666369$$

$$\Pr(2) = \frac{130}{1121} = 0.115968$$

$$\Pr(3) = \frac{124}{1121} = 0.110616$$

$$\Pr(4) = \frac{120}{1121} = 0.107047$$

Class Probability for Month 2

$$\Pr(1) = \frac{771}{1152} = 0.669271$$

$$\Pr(2) = \frac{381}{1152} = 0.330730$$

Pr(3) = 0

Pr(4) = 0

Class Probability for Month 3

$$\Pr(1) = \frac{923}{1385} = 0.666426$$

$$\Pr(2) = \frac{300}{1385} = 0.216607$$

$$\Pr(3) = \frac{78}{1385} = 0.056318$$

$$\Pr(4) \frac{84}{1385} = 0.060650$$

The results obtained, which are the predicted class labels, were compared with the extracted class labels to detect and analyze the number of occurrences of each class. The outputs for three months are shown in Tables 2, 3 and 4.

Table 2: Month 1

Class labels	**1**	**2**	**3**	**4**
1.	999041	49497	0	0
2.	3	8	0	0
3.	0	0	13	0
4.	1	0	3	7
Total no. of records	999045	49505	16	7

Table 3: Month 2

Class labels	1	2
1	493990	203018
2.	0	1
Total no. of records	493990	203019

Table 4: Month 3

Class labels	1	2	3	4
1.	947133	96124	5280	16
2.	1	11	2	1
3.	1	0	2	0
4.	0	0	0	1
Total no. of records	947135	96135	5284	18

Percentage Accuracy

The Percentage Accuracy *A%* is obtained by finding the ratio of the TRUE POSITIVE to the Total Data Set, using the formula

$$A\% = \frac{tp}{tp + tn} \quad (3)$$

$$tp = tp(1,1) + tp(2,2) + tp(3,3) + tp(4,4) \quad (4)$$

where *A%* is the Percentage Accuracy, *tp* is the true positive, that is the records that are correctly classified, and *tn* is the true negative, that is, wrongly classified.

For Month 1,

tp = 999041 + 8 +13 + 7 = 999069

total records (tp + tn) = 1048573

$$A\% = \frac{999069}{1048573} x\ 100\% = 95.28\%$$

For Month 2,

tp = 593990 + 1 = 593991

total records (tp + tn) = 697009

$$A\% = \frac{593991}{697009} x\ 100\% = 85.22\%$$

For Month 3,

tp = 947133 +11 + 2 + 1 = 947147

total records (tp + tn) = 1048572

$$A\% = \frac{947147}{1048572} x\ 100\% = 90.33\%$$

A Summary of the result is shown on Table 5.

Table 5: Percentage Accuracy of Classified Data

Class	No. of records	Correctly Classified (tp)	Wrongly Classified (tn)	% Accuracy
Month 1	1048573	999069	49504	95.28%
Month 2	697009	593991	103018	85.22%
Month 3	1048572	947147	101425	90.33%
Total	2794154	2540207	253947	90.91%

From the above table, the average accuracy of classification using the Bayesian Information Criterion model is over 90% when carried out on real life Webserver log data obtained over a period of 3 months.

CONCLUSION

From the analysis above, it has been made clear that Bayesian approach takes care of the problem of false alarms, and therefore, it is a standard tool for Intrusion Detection.

Security matters cannot be taken lightly in our present day information handling and transfer. This is compounded by man's growing dependence on Information Technology. Even Third World countries are not left out, considering the fact that up-to-date technology has been developed, at very low cost. Internet coverage has also expanded astronomically, reaching to some

very remote areas. While this has led to improved means of business transaction and information dissemination, it has also made the world very vulnerable to threats and security frauds.

Real life data already preprocessed and discretized by Onyekwelu, et al(2017) were used. The discretized data was analyzed to reduce redundancy and solve the problem of imbalanced data set. Training and testing datasets where obtained, and a Bayes classifier, based on the Bayesian Information Criterion, was implemented on a Windows 7 Professional system running on Intel(R) Core™2 Duo CPU, 2.53GHz, 4.00GB RAM, using C++ programming language. The class labels were compared from the results obtained. The results obtained are reasonable, but can be improved upon. The tools provide a means of solving the problems of data and information security.

REFERENCES

Adetunmbi, A. O. (2008): Intrusion Detection Based On Machine Learning Techniques, A Ph.D. Theses in the Department of Computer Science, Federal University of Technology, Akure, Nigeria.

Ashoor, A. S. and Gore, S. (2011): Intrusion Detection System (IDS) &Intrusion Prevention System (IPS): Case Study, in International Journal of Scientific & Engineering Research Volume 2, Issue 7, ISSN 2229-5518,https://www .researchgate .net/ publication/ 266 232983_Intrusion_ Detection_System_IDS_Case_Study

Axelsson, S. and Sands, D. (2006): Understanding Intrusion Detection Through Visualization, Springer Science+Business Media, Inc. ISBN-13: 978-0-387-27634-2.

Grace, L.K J, Maheswari, V. and Nagamalai, D (2011). Analysis of web logs and web user in web mining, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, http://arxiv.org /ftp/arxiv /papers/1101/1101. 5668.pdf

Li, W. (2012): Real Time Intrusion Detection System Using Genetic Algorithm, in IEEE Transactions On Parallel And Distributed Systems, Vol. 23, No.

Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. WIREs Computational Statistics 4, 199{203

Onyekwelu, B. A. Alese, B. K. Adetunmbi, A. O. (2017) "Pre-Processing of University Webserver Log Files for Intrusion Detection", International Journal of Computer Network and Information Security(IJCNIS), Vol.9, No.1, pp.20-30, 2017.DOI: 10.5815/ijcnis.2017.01.03

Peng, T. and Zuo, W. (2006), Data Mining for Network Intrusion Detection System in Real Time , in International Journal of Computer Science and Network Security, VOL.6 No.2B, http://paper. ijcsns.org/07_book/ 200602/ 200602 C11.pdf

Ramyachitra, D. and .Manikandan, P. (2014) Imbalanced Dataset Classification And Solutions: A Review, International Journal of Computing and Business Research (IJCBR) ISSN (Online) : 2229-6166 Volume 5 Issue 4.

Van Trees, H. L. (2001): Detection, Estimation and Modulation Theory, Part III: Radar–Sonar Signal Processing and Gaussian Signals in Noise. John Wiley & Sons, Inc. ISBNs: 0-471-10793-X (Paperback); 0-471-22109-0 (Electronic)

Vijayarani , S. and Sylviaa, M. (2015), Intrusion Detection System – A Study, in International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1, DOI : 10.5121/ijsptm. 2015. 4104 31 , http://airccse.org/journal/ijsptm/papers/4115ijsptm04.pdf