Design of Optimal Acceptance Sampling Plan for Network Intrusion Detection

C. MadhusudhanaRao

Abstract: Network Intrusion Detection Systems (NIDS) protects networks connected to the internet from malicious attacks by monitoring network flows predominantly at fragment level in network layer. Inspecting every fragment of a network flow is computationally prohibitive. The Acceptance Sampling for Network Intrusion Detection (ASNID) method avoids hundred percent inspections of fragments to detect anomalous flows. This study proposes a model to determine optimal acceptance sample size. Further, this study also proposes a model for estimating the cost of computational effort.

Key words: Acceptance Sampling, Expected Total Cost, Expected Computational Effort, Geometric Mean Accuracy Index, Network Intrusion Detection.

I. INTRODUCTION

The internet is used widely nowadays for transfer of sensitive and proprietary data. The intruders make use of the vulnerabilities of the target network to circumvent security mechanisms. Network Intrusion Detection System (NIDS) protects networks connected to the internet from malicious attacks by monitoring network flows and inspecting each fragment of a network flow. The links can hold many network flows due to availability of high capacity. Inspecting all network flows and each fragment of a network flow is computationally prohibitive.

Androulidakis and Papavassiliou in [1] used a selective flow sampling method to detect the port scan and host scan attacks using entropy and suggested that flow sampling is more appropriate for network intrusion detection. The authors of this study in [2] employed acceptance sampling for network intrusion detection to detect Land, Xmass, Nespel, Rose, Winnuke, NULL Scan, Teardrop, Fraggle, Port scan, Host scan attacks. In this method, a randomly chosen sample of fragments from a network flow is inspected for detecting whether it is anomalous or not. The metric employed for performance evaluation of the method is Geometric Mean Accuracy Index (GMAI). This method reduces the computational effort by a factor of \( 0 < k < 1 \), where \( k \) is the ratio of sample size to total fragments of a network flow.

When sample size is small, GMAI is low and computational effort is also low. When sample size increases, GMAI and computational effort increases. Large sample size is more reliable, but it is more computationally costly. A trade-off between GMAI and computational effort is to be achieved. At optimal sample size, the computational effort will attain a fair minimum value, while the GMAI is at reasonable maximum value as shown in Figure: I hereunder.

II. REVIEW OF RELATED WORK

NIDS at network layer level inspects each fragment of a network flow until a fragment is found malicious and classifies such network flow as anomalous flow. Otherwise, the network flow is classified as normal flow. Many statistical methods were proposed to detect the network intrusions. These statistical methods fit the normal behaviour of network traffic into a statistical model. Then a statistical inference test is conducted on the network traffic to determine an abnormal traffic belongs to this model.

Eskin [3] proposed a statistical model that considers an element of \( x \) of data follows majority distribution M and log likelihood C is determined. Detecting intrusion is equivalent to measure log likelihood of the test data. If log likelihood is less than C, the data is normal otherwise the data is anomalous. Mei-ling Shyu et al [4] proposed Principal Component Analysis (PCA) based Predictive model for network intrusion detection. Manikopoulos and Papavassiliou [5] introduced a method to classify network anomalous flows using statistical modelling and neural networks.

An intrusion detection method using payload information is proposed in [6]. It computes the mean frequency of each ASCII character in pay load and also the variance and standard deviation. This method computes the required statistics for each payload and this model is used during training. The payload of each incoming packet is scanned and statistics are computed during detection.

Figure 1: Sample size optimization

This study proposes a model for determining optimal sample size for accepting sampling in network intrusion detection. This study also proposes a model for estimating the cost of computational effort. The rest of the paper is organized as follows. Section II reviews the related work. Section III describes the model for determining optimal sample size for acceptance sampling. Section IV presents a model for Computing Expected Computational Effort (ECE). The simulation model for determining optimal sample size is presented in section V. Section VI gives analysis of experimental data. The conclusions of the study are reported in section VII.
This is compared against the model created during training. If there is a significant difference, the packet is declared as anomalous.

Song et al. [7] proposed a conditional anomaly detection method that obtains differences among packet header data attributes. This method computes maximum likelihood estimation to existing data that fit to the Gaussian Mixture Model. During detection data records are tested against this model.

Lu and Ghorbani [8] presented a combination of Discrete Wavelet Transform (DWT) and system identification theory method for intrusion detection. This method detects anomalous signals from normal ones. The normal traffic is given as input to the method. The output is called as residuals and these will be close to 0. If the output has lot of peaks and they are declared as anomalies.

Decision Trees(DT) such as C 4.5[9],CART[11], soft computing techniques such as fuzzy logic model[13], Artificial Neural Networks (ANN)[10], genetic algorithm[12] and rule based methods such as snort [14] are also proposed by researchers. The researchers in [15-18] used a combination of methods to increase the detection accuracy. The work reported in [3-18] considered all fragments of the network flow to detect a flow as anomalous or normal. Paul Barford et al [19] used a wavelet based filter method to select the network flows as input to detect intrusions. Nick Duffield [20] suggested that the random sampling is better than filtering as all the details retained in the sample, no latency in selecting the sample. Jianning Mai [21] proposed a port scan detection technique in which random samples of packets are inspected for distorted traffic features.

The research work reported above has not considered the trade off between accuracy in detecting intrusions and expected total cost. More over these methods are computationally prohibitive. Hence it is motivated to develop a model to determine the optimal sample size that maximizes the accuracy and minimizes computational effort and it is presented in the next sections.

III. MODEL FOR DETERMINING OPTIMAL ACCEPTANCE SAMPLE SIZE

In Acceptance Sampling for Network Intrusion Detection (ASNID) method, a sample size of ‘n’ fragments is selected randomly for inspection from a lot size of ‘N’ fragments that constitute a network flow. Obviously, the acceptance number (c) is taken as zero. The fragments in the sample are inspected until the occurrence of malicious fragment and thereof the network flow is classified as an anomalous flow. On the other hand, if it is found that none of the fragments in the sample is malicious then the network flow is classified as a normal flow. The ASNID method may lead to two types of errors. It may classify a normal fragment as malicious that is referred to as Type-I error. Similarly, classifying a malicious fragment as normal is referred to as Type-II error.

It is obvious that larger the sample size, lesser the Type-I and Type-II errors and better the GMAI. Conversely, smaller the sample size, higher the Type-I and Type-II errors and lesser GMAI. As the sample size increases, the opportunity cost of better GMAI increases as the cost of computational effort increases. However, the opportunity cost of Type-I and Type-II errors decreases. The sum of the opportunity cost of better GMAI; and the opportunity cost Type-I and Type-II errors referred to as total cost. It is a function of sample size and likely to be unimodal. As it is difficult to estimate the above said opportunity costs as a function of sample size, determining the optimal sample size for which the total cost would be minimum is a challenge. However, a novel model for determining the optimal sample size is proposed. The notations of the proposed model are given hereunder in Table: I.

| Notation | Description |
|----------|-------------|
| P_{TP}   | Probability of detecting anomalous flows as anomalous flow |
| P_{TN}   | Probability of detecting normal flows as normal flow |
| P_{FP}   | Probability of detecting anomalous flows as normal flow or equivalently probability of Type-II error |
| P_{FN}   | Probability of detecting normal flows as anomalous flow or equivalently probability of Type-I error |
| C_{TP}   | Opportunity cost of classifying anomalous flows as anomalous flow |
| C_{TN}   | Opportunity cost of classifying normal flows as normal flow |
| C_{FP}   | Opportunity cost of classifying anomalous flows as normal flow |
| C_{FN}   | Opportunity cost of classifying normal flows as anomalous flow |

The above said probabilities can be computed from the elements of the confusion matrix resulting from a given acceptance sampling plan using the equations (3.1), (3.2), (3.3) and (3.4).

\[
P_{TP} = \frac{TP}{TP + FN + FP + TN} \quad (3.1)
\]

\[
P_{TN} = \frac{TN}{TP + FN + FP + TN} \quad (3.2)
\]

\[
P_{FP} = \frac{FP}{TP + FN + FP + TN} \quad (3.3)
\]

\[
P_{FN} = \frac{FN}{TP + FN + FP + TN} \quad (3.4)
\]

The above said probabilities and opportunity costs are presented in matrix forms as shown in Table II and Table III.

| Input Flows | Detected Flows | Anomalous | Normal |
|-------------|----------------|-----------|--------|
|             | P_{TP}         | P_{TN}    |        |

| Input Flows | Detected Flows | Anomalous | Normal |
|-------------|----------------|-----------|--------|
|             | C_{TP}         | C_{TN}    |        |

| Input Flows | Detected Flows | Anomalous | Normal |
|-------------|----------------|-----------|--------|
|             | C_{TP}         | C_{TN}    |        |

The Expected Total Cost (ETC) of ASNID for given acceptance sampling plan is derived as Equation (3.5).

\[
ETC = C_{TP} * P_{TP} + C_{TN} * P_{TN} + C_{FP} * P_{FP} + C_{FN} * P_{FN} \quad (3.5)
\]

Obviously, the opportunity costs classifying anomalous flows as anomalous (C_{TP}) and normal flows as normal (C_{TN}) are zero.
Therefore, the Equation (3.5) can be reduced to Equation (3.6).

\[
ETC = C_{FN} * P_{FP} + C_{FP} * P_{FP}
\]

(3.6)
The sample size, which minimizes objective function given as Equation (3.7) is optimal.

Minimize \(ETC = C_{FN} * P_{FN} + C_{FP} * P_{FP}\) \hspace{1cm} (3.7)

Intuitively, in the above equation, \(C_{FP}\) is significantly less than that of \(C_{FN}\). Hence, the ETC can be derived without loss of generality as given in Equation (3.8). Therefore Minimizing \((ETC)\) is minimizing \((C_{FN}/(C_{FN}+C_{FP}))\).

Minimize \(ETC = C_{FN} / (C_{FN} + C_{FP}) = (P_{FN} / (P_{FN}+P_{FP}))\) \hspace{1cm} (3.8)

Generally, it is difficult to estimate the costs, \(C_{FN}\) and \(C_{FP}\). Whenever it is difficult to estimates the costs, service level is taken as surrogate. Here, the GMAI can be taken as surrogate for ETC. A security analyst can determine the optimal sample size based the assumption of probability of anomalous flows and service level requirement through simulation. The model for computing Expected Computational Effort (ECE) and experimentation through simulation is presented in the next section.

IV. MODEL FOR COMPUTING EXPECTED COMPUTATIONAL EFFORT (ECE)

A network flow is partitioned into a number of packets of size 65535 bytes each. Each packet is partitioned into a number of fragments of size 1500 bytes or part each.

Number of packets per a network flow is

\[
NoP = \frac{NFS}{65535}
\]

(4.1)

Where NFS is network flow size in bytes

Number of fragments for the network flow under consideration is

\[
NoF = \sum_{i=1}^{r} \left( \frac{Pct_i}{1500} \right)
\]

(4.2)

Where \(Pct_i\) is \(i^{th}\) Packet Size.

\(p = \text{percentage of malicious fragments in } NoF\)

\(n = \text{sample size}\)

\(r = \text{np, number of malicious fragments in random sample}\)

Axiom 1:

If \(i\) fragments are not malicious and fragment \(i\), is malicious then the number of ways \((r-i)\) malicious fragments occur is

\(w_i = (n-i)C_{r-(n-1)}\) for \(1 \leq i \leq (n-r+1)\).

Axiom 2:

The probability of that \((i-1)\) fragments are not malicious and fragment \(i\) is malicious, is \(P = \frac{w_i}{n} \) for \(1 \leq i \leq (n-r+1), zero otherwise and \(\sum_{i} P_i = 1\).

Accordingly, the computational effort in terms of the number of fragments to be inspected for the best case, worst case and average case are shown in Table IV.

V. SIMULATION MODEL

The simulation model for determining optimal sample size is presented in this section. The synthetic network flow data set generated using the model presented in [22] is considered. The Cardinality of a synthetic network flow data set is fixed as 10,000 and the number of fragments in a network flow varies with the flow size. Ninety simulations are conducted based on the factors and levels shown in Table V.

| Table IV: Factors and Levels |
|-----------------------------|
| Factor | Number of Levels |
| Sample Size (Percentage of fragments of a network flow under consideration) | 10 | (Varying 10% to 100% in steps of 10%) |
| Percentage of Anomalous Flows in a set of Network Flows | 9 | (Varying 10% to 90% in steps of 10%) |

The elements of confusion matrix are generated for each of ninety simulations. The corresponding \(P_{FP}, P_{FN}\) and ETC are computed and tabulated in Table VI, Table VII and Table VIII. Its graphical representation is shown in Figure 2 and Figure 3. The analysis of experimental data is presented in the next section.

VI. EXPERIMENTAL RESULTS AND OUTPUT ANALYSIS

It is evident from the Table V, Table VI and Figure 2 that as sample size increases the GMAI increases monotonically until it attains the value of one ETC decreases until it attains the value of zero. Further, it is noticed that GMAI attains the value of one ETC attains the value of zero at 60% of sample size for all percentages of anomalous flows. Hence, the optimal sample size is 60% as it yields GMAI ETC one and zero respectively. Further, the optimal sample size is 40% for percentages of anomalous flows 50% and beyond as depicted in Figure 3. The observations made with respect to ECE are as follows:

1. The ECE for inspecting the optimal sample size of 60% is reduced by a factor of 0.4.
2. The ECE increases as sample size increases.
3. The ECE decreases as the percentage of

Figure 2: Sample Size Vs GMAI & ETC for 50% anomalous flows

Figure 3: Percentage of Anomalous Flows Vs Sample Size
Design of Optimal Acceptance Sampling Plan for Network Intrusion Detection

malicious fragments increases.

VII. CONCLUSIONS

This study proposed mathematical models for computing ETC and ECE given the sample size and percentage of anomalous flows. Further, a simulation model is developed to determine optimal sample size utilizing ASNID method. A security analyst can employ these models for determining the optimal sample size based on the assumption probability of anomalous flows and service level requirement through simulation. It facilitates the analyst to choose an optimal sample size on his own perception of security policy requirements by specifying the corresponding GMAI and ECE.

Table V: GMAI & ETC for the Percentage of Anomalous Flows up to 40

| % Anomalous Flows | Sample size in % | GMAI | ETC | GMAI | ETC | GMAI | ETC |
|-------------------|-----------------|------|-----|------|-----|------|-----|
|                   |                 | 10   | 20  | 30   | 40  | 50   | 60  |
|                   |                 | 0.995816 | 0.997387 | 0.998431 | 0.998956 | 0.999478 | 1 |
|                   |                 | 1     |     | 1    |     | 1    |     |
|                   |                 | 0.996154 | 0.998464 | 0.999232 | 0.999488 | 0.999974 | 0 |
|                   |                 | 1     |     | 1    |     | 1    |     |

Table VI: GMAI & ETC for the Percentage of Anomalous Flows from 50 to 90

| % Anomalous Flows | Sample size in % | GMAI | ETC | GMAI | ETC | GMAI | ETC |
|-------------------|-----------------|------|-----|------|-----|------|-----|
|                   |                 | 50   | 60  | 70   | 80  | 90   |
|                   |                 | 0.998487 | 0.999487 | 0.999526 | 0.999548 | 0.999571 | 1 |
|                   |                 | 1     |     | 1    |     | 1    |     |
|                   |                 | 0.998921 | 0.999668 | 0.999879 | 0.999927 | 1 |
|                   |                 | 1     |     | 1    |     | 1    |     |

Table VII: Expected Computational Effort

| Sample Size in % | Percentage of Malicious Fragments |
|------------------|----------------------------------|
|                  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| 10               | 5.5 | 3.666667 | 2.75 | 2.2 | 1.833333 | 1.571429 | 1.375 | 1.222222 | 1.1 |
| 20               | 10.5 | 4.2 | 3 | 2.333333 | 1.900901 | 1.615385 | 1.4 | 1.235294 | 1.105263 |
| 30               | 15.5 | 4.28571 | 3.1 | 2.384615 | 1.9375 | 1.631579 | 1.409091 | 1.24 | 1.107143 |
| 40               | 20.5 | 4.555556 | 3.153846 | 2.411765 | 1.952381 | 1.64 | 1.413793 | 1.242422 | 1.108108 |
| 50               | 25.5 | 4.636364 | 3.1875 | 2.428571 | 1.961538 | 1.645161 | 1.416667 | 1.243902 | 1.108696 |
| 60               | 30.5 | 4.692308 | 3.210526 | 2.44 | 1.967742 | 1.648649 | 1.418605 | 1.244898 | 1.109091 |
| 70               | 35.5 | 4.733333 | 3.227273 | 2.448276 | 1.972222 | 1.651163 | 1.42 | 1.245614 | 1.109375 |
| 80               | 40.5 | 4.764706 | 3.24 | 2.454545 | 1.97561 | 1.653061 | 1.421053 | 1.246154 | 1.109589 |
| 90               | 45.5 | 4.789474 | 3.25 | 2.459459 | 1.978261 | 1.654545 | 1.421875 | 1.246575 | 1.109756 |
| 100              | 50.5 | 4.809524 | 3.258065 | 2.463415 | 1.980392 | 1.655738 | 1.422355 | 1.246914 | 1.10989 |
ACKNOWLEDGMENT

The author of this research paper is gratefully acknowledged Dr. M.M. Naidu, Professor of CSE (Retd), SVU College of Engineering, Tirupati for his advice in doing this research work.

REFERENCES

1. Georgios Androulidakis, Vassilis Chatzigiannakis, Syeem Papavassiliou, “Network Anomaly Detection and Classification via Opportunistic Sampling”, IEEE Networks, Vol 23, NO.9, 2009, pp 6-12.

2. C. Madhusudhanarao, Dr. M.M. Naidu, “Acceptance Sampling for Network Intrusion Detection” Journal of Theoretical and Applied Information Technology Vol 95, NO24, pp 6707-6718, December 2017.

3. Eleazar Eskin, “Anomaly Detection over Noisy data using learned probability distributions” in Proc. 7th International Conference on Machine Learning. Morgan Kaufmann, 2000, pp. 255–262.

4. M L Shyu, S C Chen, K Sarinna pakorn, and L Chang, “A novel anomaly detection scheme based on principal component classifier,” IEEE Foundations and New Directions of Data Mining Workshop, pp. 172–179, Nov. 2003.

5. C. Manikopoulos, S. Papavassiliou, “Network Intrusion and Fault Detection: A Statistical Anomaly Approach,” IEEE Commun. Mag., vol. 40, no. 10, pp. 76–82, October 2002.

6. K. Wang and S. J. Stolfo, “Anomalous Payload-Based Network Intrusion Detection,” in Proceedings of Recent Advances in Intrusion Detection. Springer, 2004, pp. 203–222.

7. X Song, M Wu, C Jermaine, and S Ranka, “Conditional Anomaly Detection,” IEEE Transactions on Knowledge and Data Eng., vol. 19, pp. 631–645, 2007.

8. W Lu and A A Ghorbani, “Network Anomaly Detection Based on Wavelet Analysis,” EURASIP Journal on Advances in Signal Processing, vol. 2009, no. 837601, January 2009.

9. Sandya Peddabachigari, Ajith Abraham, Crina Grosan, Johnson Thomas, “ Modeling Intrusion Detection System using Hybrid Intelligent Systems” Journal of Network and Computer Applications, Vol 30, No.1, 2007, pp 114-132.

10. Srinivas Makkamala, Andrew H. Sung “Feature Selection for Intrusion Detection using Neural Networks and Support Vector Machines” Journal of transportation research board, Vol 1822, 2003,pp1822-5.

11. Srilatha Chebrolu, Ajith Abraham, Johnson P. Thomas, “Feature deduction and ensemble design of intrusion detection systems” Journal of Computers & Security, Vol 24, 2005,pp 295-307.

12. Taeshik Shon, Yongdoo Kim, Cheolwon Lee, Jon sub Moon, “A Machine Learning Framework for Network Anomaly Detection using SVM and GA” Proceedings of the IEEE Workshop on Information Assurance and Security, United States Military Academy, West Point, NY,June.15-19, 2005,pp 176-183.

13. 13.Surat Srinoy, Wereas Kurutach, Witcha Chimphlee, Sirporn Chimphlee, Santi Sounsri, “Computer Intrusion Detection with Clustering and Anomaly Detection Using ICA and Rough, Fuzzy” Proceedings of the 4th WSEAS Int. Conference on Information Security, Communications and Computers, Tenerife, Spain, December, 16-18, 2005, pp 252-257.

14. M. Roesch, “Snort-Lightweight Intrusion Detection for Networks,” Proc. 13th USENIX Conference on System Administration, Seattle, Washington, November, 7-12, 1999, pp 229–238.

15. 15. S Jiang, X Song, H Wang, J J Han, Q H. Li, “A Clustering Based Method for Unsupervised Intrusion Detections,” Pattern Recognition Letters, Vol. 27, no. 7, 2006, pp. 802–810.

16. Shekhar R. Gaddam, Vir V. Phoha, Kiran S. Balagani,“K-Mean+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods”, IEEE Transactions on knowledge and data Engineering , Vol.19,No.3, 2007, pp 1-10.

17. Anamthar Prabakar Muniyandi, R. Rajeswari, R. Rajaram “Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm” Procedia Engineering, Vol.30, 2012. pp 174 – 182.

18. Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, “A K-means and naive bayes learning approach for better intrusion detection,” Information Technology Journal., Vol. 10, No. 3, 2011, pp 648-655.

19. P Barford, j Kline, D. Plonka, and A. Ron “A Signal Analysis of Network Traffic Anomalies”, In Proc. ACM SIGCOMM IMW’02, pages 71–82, Marseille, France. Nov. 2002.

20. N Duffield “Sampling for Passive Internet Measurement: A Review”. Statistical Science, Vol. 19, No.3, pp 472–498, 2004.

21. J Mai, A Sridharan, C N Chuah, T Ye, and H Zang “Impact of Packet Sampling on Port scan Detection” Technical Report RR06-ATL-043166, Sprint ATL, 2006.

22. C. Madhusudhanaraao, M. M. Naidu , “A Model for Generating Synthetic Network flows and Accuracy Index for Evaluation of Anomaly Network Intrusion Detection Systems”, Indian Journal of Science and Technology, Vol 10 No.14, pp 1-16, 2017.

AUTHORS PROFILE

Dr. C.Madhusudhanarao is currently working as Professor of Computer Science and Systems Engineering in Sree Vidyanikethan Engineering College, Tirupati. He received B.Tech degree from Bapatla Engineering College, M.Tech from P.I.D.A College of Engineering, Kalaburgi, and Ph.D from SVU College of Engineering, Tirupati. Currently, he is coordinating the research activity of Centre for Computer Networks and Cyber security. His Research interests are Cyber Security, Blockchain Technology and Machine learning.