International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 13 (2018) pp. 11030-11034

Flow Sampling for Network Intrusion Detection –An Acceptance Sampling Approach

C. MadhusudhanaRao
Research Scholar,
Department of Computer Science and Engineering,
Sri VenkateswaraUniversity College of Engineering,
Tirupati -517502, India

Dr. M.M.Naidu
Professor (Retired)
Department of Computer Science and Engineering,
Sri VenkateswaraUniversity College of Engineering,
Tirupati -517502, India

Abstract

A Network Intrusion Detection System (NIDS) detects anomalous network flows by inspecting the incoming fragments for malicious patterns and notifies to the network administrator, if any. Hundred percent inspections of all the fragments of network flows for detecting malicious fragments and thereof anomalous flows are highly prohibitive. The Acceptance Sampling for Network Intrusion Detection Method (ASNID) method inspects a randomly chosen sample of fragments from a network flow for detecting whether it is anomalous or not. It 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. This study proposes a novel flow sampling method utilizing acceptance sampling for network intrusion detection and this method further reduces the computational effort. It is proved experimentally that the Geometric Mean Accuracy Index (GMAI), the performance metric monotonically increases as percentage of anomalous network flows increases. The GMAI monotonically increases as the sample of percentage of network flows increases. The computational effort is further reduced by a factor of $0 <j<1$ where j is the ratio of sample of flows to total number of network flows.

Keywords: Acceptance Sampling, Flow sampling, Geometric Mean Accuracy Index, Network Work Intrusion Detection.

INTRODUCTION

NIDS protects networks connected to the internet from network attacks by monitoring network flows predominantly at fragment level. The methods proposed in the past, inspect each fragment of a network flow for detecting whether a fragment not malicious by comparing with ominous fragment patterns. Obviously, the network flow is anomalous if at least one fragment is detected as malicious.

The volume of network flows has been steadily increasing in the recent years. The increase in volume of network flows on high capacity links leads to high computational effort for intrusion detection methods [1]. Sampling techniques have been widely used to deal with the growing network flows [2]. Various approaches and methods were suggested to sample the network flows that preserve the characteristics of the monitored traffic such as the number of transferred bytes, packets and flows [3].

The use of sampled data for intrusion detection is problematic [4], as any sampling necessarily impacts the effectiveness of the intrusion detection and data analysis algorithms. These algorithms are based on pattern recognition and statistical traffic analysis, and the distortion of traffic features can significantly increase the error rate of these underlying methods. The loss of information introduced by sampling methods also negatively impacts any potential detection. This challenge motivated to propose Acceptance Sampling for Network Intrusion Detection Method (ASNID) method [5] and this study proposes network flow sampling utilizing acceptance sample for network intrusion detection. The performance metric employed for its performance evaluation is Geometric Mean Accuracy Index (GMAI) [6]. It is experimentally through simulation proved that the proposed system maximizes GMAI while minimizing computational effort.

The rest of the paper is structured as follows. Section 2 reviews the related work. Section 3 proposes the adaption of acceptance sampling for Network Intrusion Detection. The performance evaluation of proposed method through simulation is presented in section 4. Section 5 presents analysis of experimental data. The contributions of the study and scope for future study are reported in section 6.

REVIEW OF RELATED WORK

Dorothy E. Denning [7] proposed an Intrusion Detection System (IDS) that is able to identify attacks against vulnerable services and applications, privilege violations, unauthorized logins and access to sensitive files. Network intrusion detection based on naïve Bayesian classifier (NB) [8], support vector machine (SVM)[9] have been proposed by researchers to classify the network flows as normal or anomaly. Yihua Liao et al [10] proposed k-nearest neighbour (KNN) classifier to classify TCP/IP sessions as normal or anomaly on DARPA BSD system call data.

Decision Trees(DT) such as C 4.5[13],CART[12], soft computing techniques such as fuzzy logic model[8], Artificial Neural Networks (ANN)[11], genetic algorithm[14] and rule based methods such as snort [15] are also proposed by researchers.

A method that cascade k-Means clustering and the ID3 decision tree for classifying anomalous and normal activities is proposed in [16]. Similarly cascading k-means clustering and C4.5 [17], cascading k-means clustering and naïve Bayesian classifier [18] is proposed to improve the detection accuracy.

An application of packet sampling for traffic analysis has been studied by Duffield *et al.* [19] and Estan *et al.* [20]. The work by Ali *et al.* [21] proposed a new type of packet sampling combined with an anomaly detector, where anomalous packets were sampled with higher probability.

In the case of *flow sampling*, monitored traffic is aggregated into network flows, and the sampling is applied not to the particular packets but to the whole flows. The main benefit is better accuracy when compared with packet sampling [22], but this requires more memory and CPU power. A comparison of packet and flow sampling can be found in the work of Hohn and Veitch [22]. Based on the results, flow sampling is superior in the preservation of the flow distributions. A comprehensive literature review can be found in the work of Duffield [23].

Mai *et al.* [4] evaluated how the performance of anomaly detection algorithms is affected by the selected sampling methods. The work by Androulidakis *et al.* [24] proposed a selective flow sampling suitable for anomaly detection. The authors of this study proved that random sampling proves better than selective sampling in [5]. All the above studies not measured the effectiveness of flow sampling in the detection of intrusions. The next section presents Acceptance sampling for network intrusion detection.

ACCEPTANCE SAMPLING FOR NETWORK INTRUSION DETECTION

In acceptance sampling, a sample of size 'n' chosen randomly from lot of size 'N' is inspected. A lot is accepted provided the number of defectives are less than or equal to c, the acceptance number. The proportion of defectives acceptable to a customer is known Acceptable Quality Level (AQL). Similarly, the proportion of defectives not acceptable to a customer is known as Lot Tolerance Percent Defective (LTPD). It is obvious that in acceptance sampling, the probability of rejecting a lot with the proportion of defectives not exceeding AQL is not zero.

Similarly, the probability of acceptance of a lot with proportion of defectives exceeding LTPD is not zero. The first possibility is referred to as producer's risk (α) whereas the second possibility is referred to as consumers' risk (β). The producer's risk and consumer's risk are also known as Type-I and Type-II errors respectively. A sampling plan (n, c) is designed with the objective of minimizing α and β, given N, AQL and LTPD. It is obvious that 100% inspection of network flow may guarantee 100 % accuracy in detecting anomalous flows. However, the highly prohibitive computational effort affects application response time adversely.

The parameters of acceptance sampling in statistical quality control and network intrusion detection are given in Table 1.

Table 1: Parameters of Acceptance Sampling in Statistical Quality Control and Network Intrusion Detection

Parameter	Notation	Statistical Quality Control	Network Intrusion Detection
Lot size	N	Number of units in a lot submitted for inspection	Number of network flows
sample size	n	Number of randomly chosen units for inspection	Number of randomly chosen network flows for inspection
acceptance number	c	Threshold of defective units in a sample for accepting a lot, usually nonzero	Threshold of anomalous flows in a sample
% defective / anomalous	p	Percent defectives in a lot submitted for inspection	Percent anomalous network flows submitted for inspection

The steps of overall logic of acceptance sampling for detecting anomalous flows are as follows:

1. Initially choose a sample of network flows n randomly from a set of network flows
2. Inspect the first network flow
3. Assume that the network flow under consideration as a normal flow initially
4. Choose a random sample of fragments randomly from the network flow
5. Inspect the first / next fragment
6. If the fragment is detected as malicious then classify the flow as anomalous and stop
7. Otherwise go to step 5 while the next fragment exists else inspect next network flow

The performance analysis and experimentation is explained in the next section.

Table 2: Percentage of network flows Vs Percentage of Anomalous Flows at 40% Sample size

Sample of flows in %	10	20	30	40	50	60	70	80	90	100
%Anomalous										
10	0.998806	0.999502	0.999673	0.999709	0.99973	0.999751	0.999755	0.999755	0.999759	0.999786
20	0.999823	0.999832	0.999834	0.999846	0.999857	0.999858	0.999864	0.99987	0.999877	0.999879
30	0.999888	0.99989	0.999891	0.999895	0.999899	0.999899	0.9999	0.999901	0.999901	0.999904
40	0.999906	0.999911	0.999915	0.999928	0.999929	0.999938	0.999945	0.999962	1	1
50	0.999941	0.999954	0.999973	0.999986	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1	1
70	1	1	1	1	1	1	1	1	1	1
80	1	1	1	1	1	1	1	1	1	1
90	1	1	1	1	1	1	1	1	1	1

PERFORMANCE EVALUATION

This section presents performance evaluation of flow sampling for network intrusion detection utilizing acceptance sampling. For each synthetic network flow set of cardinality 10,000, simulation is performed to produce confusion matrix. The simulations are performed taking a synthetic network flow set consisting of a specific percentage of anomalous flows as input and varying the sample size in the range of [10%-100%]. Further, the percentage of anomalous flows in a network flow set is varied between 10% and 90% in steps of 10%. Nine hundred experiments are conducted, GMAI is computed for each experiment and the outcome of the experiments for 40% sample size is shown in the Table 2. Its graphical representation is shown in Figure 1 and Figure 2. The analysis of experimental data is presented in the next section.

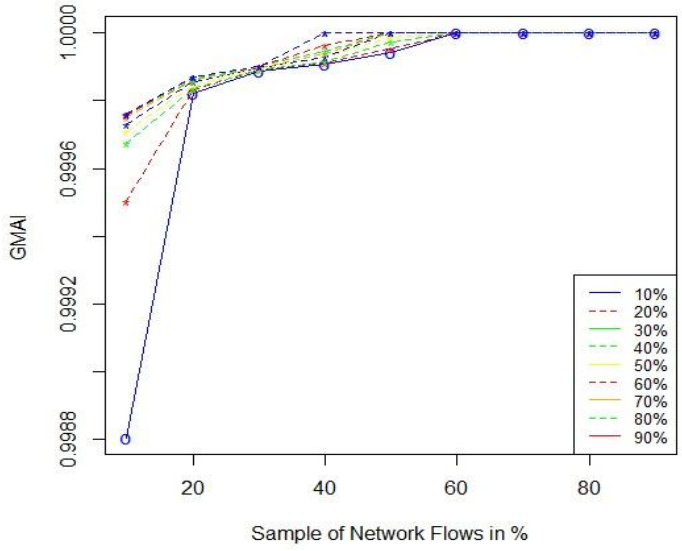

Figure 1. Sample of Network Flows in % Vs GMAI

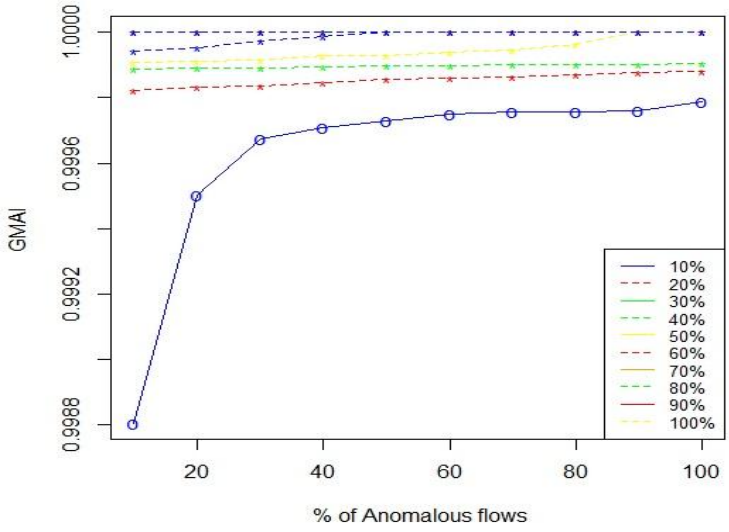

Figure 2. Percentage of Anomalous flows Vs GMAI

ANALYSIS OF EXPERIMENTAL DATA

It is evident from Figure 1 that for a given network flow associated with a specific percentage of anomalous flows, as the percentage of sample of network flows increases, the GMAI monotonically increases to one. From Figure 2, it is obvious that for a given sample size, as the percentage of anomalous flows of a set of network flows increases the GMAI monotonically increases to one. It is also observed that

1. The GMAI monotonically increases as the sample size of fragments of a network flow increases
2. The GMAI monotonically increases as the percentage of malicious fragments of a network flow increases.
3. Irrespective of percentage of anomalous flows the GMAI is one for a sample of 10% flows or more.

The time complexity of hundred percent inspections is O (N), For ASNID, it is k*O (N) where 0<k<1, the ratio of acceptance sample size to total number of fragments of a flow. For this flow sampling utilizing acceptance sampling method the time complexity is j*k* O (N).

CONCLUSIONS

The Acceptance Sampling for Network Intrusion Detection (ASNID) method avoids hundred percent inspections of fragments and thereof network flows. It is applicable to network flows of large number of fragments. It is proved experimentally that as percentage of sample size increases GMAI increases. Further, it is also experimentally proved that the GMAI is monotonically increases to one as the percentage of anomalous flows increases, for a given sample size. The GMAI monotonically increases as the percentage of malicious fragments of a network flow increases. Irrespective of percentage of anomalous flows the GMAI is one for a sample of 10% flows or more. The time complexity of hundred percent inspections is O (N), For ASNID, it is k*O (N) where 0<k<1, the ratio of acceptance sample size to total number of fragments of a flow. For this flow sampling utilizing acceptance sampling method, the computational effort is further reduced by a factor of 0 <j<1 where j is the ratio of sample of flows to total number of network flows. It is worth to investigate the optimal size of fragments of a flow and also the optimal number of network flows for inspection in future.

REFERENCES

[1] Jusko J, Rehak M., 2014, "Identifying peer-to-peer communities in the network by connection graph analysis" *International Journal of Network Management*; 24(4), pp 235–252.

[2] Carela Espaol V, Barlet Ros P, Cabellos-Aparicio A, Sol-Pareta J, 2011," Analysis of the impact of sampling on Netflow traffic classification". *Computer Networks*; 55(5), pp 1083–1099.

[3] Ali S, Haq IU, Rizvi S, Rasheed N, Sarfraz U, Mirza F, 2010 ,"On mitigating sampling-induced accuracy loss in traffic anomaly detection systems", *Computer Communication Review*, 40, pp.4–16.

[4] Mai J, Chuah C-N, Sridharan A, Ye T, Zang H, 2006, "Is sampled data sufficient for anomaly detection?" In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, IMC '06 ACM: New York, NY, USA, pp.165–176.

[5] C.MadhusudhanaRao, Dr. M.M.Naidu, 2017, "Acceptance Sampling for Network Intrusion Detection" Journal of Theoretical and Applied Information Technology, 95(24), pp 6707-6718.

[6] C.Madhusudhanarao, M M Naidu, 2017, "A model for generating synthetic network flows and Accuracy index for evaluation of anomaly network intrusion detection systems*" Indian Journal of Science and Technology*, 10(14), pp 1-16.

[7] Dorothy E. Denning, 1987, "An Intrusion-Detection Model", IEEE Transactions on Software Engineering, -13(2), pp. 222-232.

[8] Christopher Kruegel, Darren Mutz, William Robertson ,Fredrik Valeur, 2003, " Bayesian Event Classification for Intrusion Detection" *Proceedings of 19th Annual Computer Security Applications Conference*, Las Vegas, USA, pp.14-23.

[9] S. Mukkamala, G. I. Janoski, A. H. Sung, 2002, "Intrusion Detection Using Support Vector Machines", *Proceedings of the High Performance Computing Symposium - HPC*, San Diego, pp 178-183.

[10] Yihua Liao, Rao Vemuri, 2002, "Use of K-Nearest Neighbour classifier for Intrusion Detection", J*ournal of Computers & Security*, 21(5), pp 439-448

[11] Srinivas Mukkamala, Andrew H. Sung, 2003, "Feature Selection for Intrusion Detection using Neural Networks and Support Vector Machines" *Journal of transportation research board*, 1822, pp1822-1835.

[12] Srilatha Chebrolu, Ajith Abraham, Johnson P. Thomas, 2005, "Feature deduction and ensemble design of intrusion detection systems" *Journal of Computers & Security*, 24, pp 295-307

[13] Sandya Peddabachigari, Ajith Abraham, Crina Grosan, Johnson Thomas, 2007, " Modeling Intrusion Detection System using Hybrid Intelligent Systems" *Journal of Network and Computer Applications*, Vol 30(1),pp 114-132

[14] Taeshik Shon, Yongdue Kim, Cheolwon Lee, Jon sub Moon, 2005 ,"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, pp 176-183.

[15] M. Roesch, 1999, "Snort - Lightweight Intrusion Detection for Networks," *Proc. 13th USENIX Conference on System Administration*, Seattle, Washington, pp 229–238.

[16] Shekhar R. Gaddam, Vir V. Phoha, Kiran S. Balagani, 2007,"K-Means+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* ,19(3), pp 1-10.

[17] Amuthan Prabakar Muniyandi, R. Rajeswari, R. Rajaram, 2012, "Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm" *Procedia Engineering*,30,pp 174 – 182.

[18] Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, 2011,"A K-means and naive bayes learning approach for better intrusion detection," *Information Technology Journal*,10 (3), pp 648–655.

[19] Duffield N, Lund C, Thorup M, 2005, "Estimating flow distributions from sampled flow statistics", *IEEE/ACM Transactions on Networking*; 13, pp 933–946.

[20] Estan C, Keys K, Moore D, Varghese G, 2004, "Building a better Netflow", *Computer Communication Review*; 34, pp 245–256.

[21] Ali S, HaqIU, RizviS, Rasheed N, Khayam SA, Mirza F, 2010 ," On mitigating sampling-induced accuracy loss in traffic anomaly detection systems" *Computer Communication Review*; 40,pp 4–16.

[22] Hohn N, Veitch D, 2006 "Inverting sampled traffic", *IEEE Transactions on Networking*; 14(1), pp 68–80.

[23] Duffield N, 2004, "Sampling for passive internet measurement a review", *Statistical Science*, 19, pp 472–498

[24] Androulidakis G, Papavassiliou S, 2008, "Improving network anomaly detection via selective flow-based sampling", *Communications, IET*; 2(3), pp 399–409