Journal of Network and Computer Applications 87 (2017) 185–192

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling☆

W. Haider[a], J. Hu[a,*,1], J. Slay[a], B.P. Turnbull[a], Y. Xie[b]

[a] School of Engineering and Information Technology, University of New South Wales at Australian Defence Force Academy, Canberra, Australia
[b] School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, PR China

ARTICLE INFO

Keywords:
IDS
IDS dataset
Dataset evaluation
Dataset realism
Fuzzy logic
HIDS
NIDS

ABSTRACT

Prior to deploying any intrusion detection system, it is essential to obtain a realistic evaluation of its performance. However, the major problems currently faced by the research community is the lack of availability of any realistic evaluation dataset and systematic metric for assessing the quantified quality of realism of any intrusion detection system dataset. It is difficult to access and collect data from real-world enterprise networks due to business continuity and integrity issues. In response to this, in this paper, firstly, a metric using a fuzzy logic system based on the Sugeno fuzzy inference model for evaluating the quality of the realism of existing intrusion detection system datasets is proposed. Secondly, based on the proposed metric results, a synthetically realistic next generation intrusion detection systems dataset is designed and generated, and a preliminary analysis conducted to assist in the design of future intrusion detection systems. This generated dataset consists of both normal and abnormal reflections of current network activities occurring at critical cyber infrastructure levels in various enterprises. Finally, using the proposed metric, the generated dataset is analyzed to assess the quality of its realism, with its comparison with publicly available intrusion detection system datasets for verifying its superiority.

1. Introduction

Since the beginning of the 1970s, intrusion detection systems (IDSs) have been extensively adopted to protect computer networks against both known and unknown attacks (Ahmed et al., 2016; Hu et al., 2011; Haider et al., 2015). A realistic audit dataset of computer networks plays the role of evaluating the performance (e.g. accuracy and computational time) of an IDS design (Hoang et al., 2009; Haider et al., 2016). DARPA KDD98, which was generated approximately eighteen years ago and is considered the gold standard of datasets, has become outdated in terms of real-world networks normal traffic and attack behaviors (Haider et al., 2015; Creech, 2014). Although several related reports (Zuech et al., 2015; Shiravi et al., 2012; Gogoi et al., 2012; McHugh, 2000; Vasudevan et al., 2011; Sangster et al., 2009; Song et al., 2011) have attempted to extend and replace this dataset, but they have been unsuccessful because: (i) due to concerns regarding their integrity, cost and interruption of business operations, enterprises have not allowed IDS researchers to use their actual production networks for the purpose of penetration testing; and (ii) it is practically impossible for the low-resource-enabled testbeds or home-made honeypots used by existing IDS dataset generators to simulate normal traffic dynamics and possible attacks of real-world networks. Moreover, the quantified quality of realism of any IDS dataset cannot be assessed using the metric given in Shiravi et al. (2012) which is based on YES/NO arguments regarding the incompletely defined design features of an IDS dataset generation process.

To fill these gaps, in Section 2, we propose a metric based on a fuzzy logic system (FLS) that provides a theory for evaluating the quality of realism of any IDS dataset. Then, it is used to evaluate existing IDS datasets to demonstrate its capability to quantify their quality of realism. In Section 3, current IDS datasets and their quality evaluation criteria are briefly discussed to justify the need for the proposed metric and IDS dataset. Based on the results for the quality of realism evaluation obtained by the proposed metric, in Section 4, we briefly describe how the authors designed and generated a synthetically realistic IDS dataset, i.e., the next-generation IDS dataset (NGIDS-DS) using IXIA *Perfect Storm* (Penetration Tester, 2016) in conjunction with a range of commercial cyber-security-test hardware platforms. Also, the quantity of captured data and preliminary analysis of the simulation traffic are provided. In Section 5, firstly, we apply the proposed metric on the NGIDS-DS to calculate its quality of realism and compare it with those of existing datasets which proves that it is superior.

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.
E-mail address: J.hu@adfa.edu.au (J. Hu).
[1] IEEE Senior member.

http://dx.doi.org/10.1016/j.jnca.2017.03.018
Received 2 November 2016; Received in revised form 6 March 2017; Accepted 20 March 2017
Available online 28 March 2017

Secondly, the IDS dataset quality evaluation metric in Shiravi et al. (2012) is also applied on the proposed and existing IDS datasets to illustrate the difference between it and the proposed metric.

2. Evaluation of realism of IDS datasets

The need for the NGIDS-DS is justified through an IDS dataset quality of realism evaluation metric modeled using a FLS based on the Sugeno fuzzy inference engine (Sugeno and Yasukawa, 1993) which consists of four main parts: a fuzzifier; rules; an inference engine; and a defuzzifier. From the perspective of qualitative modeling, it is considered that fuzzification is the process of quantifying a qualitative subject, in this case, the quality of realism of an IDS dataset, by defining its components, i.e., an inference engine, crisp sets of input data, rules, membership functions, fuzzy linguistic variables, a fuzzy set and fuzzy linguistic terms.

The Sugeno fuzzy inference model, which is adopted as the FLS's inference engine, requires two distinct crisp input sets and the relationship(s) between their elements to form rules. Accordingly, we define these sets as $X = \{x_1, x_2, x_3 \ldots x_6\}$ and $Y = \{y_1, y_2\}$, and their input membership functions as $F_1(x_k)$ and $F_2(y_l)$ in Tables 1, 2 respectively. Set X represents the factors of a possible realistic IDS dataset acquired in chunks from related efforts (Creech and Hu, 2013; Zuech et al., 2015; Shiravi et al., 2012; Gogoi et al., 2012; McHugh, 2000; Vasudevan et al., 2011; Sangster et al., 2009; Song et al., 2011) and set Y the environmental variable for generating an IDS dataset (Davis and Magrath, 2013). Importantly, in Table 1, for factor x_4, the operational timing refers to the inclusion in the IDS dataset of peak hours, after hours, nights, weekends and time zone differences, and the industry complexity refers the varieties of cyber-based enterprises, such as communications, the military, banks, academia, health, social media, e-commerce as well as number of users and applications. Further, there are two reasons to propose and define the elements of set X and Y: (i) to establish the minimal ingredients to generate an IDS dataset; and (ii) to select and evaluate the quality of realism of any IDS dataset before applying to an IDS design.

The task of the membership function $F_1(x_k)$ is to assign a predefined output a singleton value, i.e., 0.16, where the value of k is dependent on the particular dataset under observation and the factor(s) from set X in it. It is determined by the fact that, if a dataset has maximum realism, the probability of realism is 1. Furthermore, if we express the maximum realism as 1 with respect to the contributions of the six predefined factors, for each member of set X, its share in contributing to realism will be 1/6. On the other hand, the purpose of the membership function $F_2(y_l)$ is to assign a predefined output a singleton value, i.e., 1 for real or 0.5 for synthetic to input y_l, where l is dependent on the particular dataset under observation and its type of generation environment. It actually embeds an approximation of the fact into the rule inference, i.e., the probability of the quality of realism of an IDS dataset generated over real networks will be considered the maximum and half over the synthetic ones. Moreover, the predefined assignment values (e.g., 0.16, 1 and 0.5) of the membership functions

Table 1
Crisp input set X.

Elements	Description	$F_1(x_k)$
x_1	Complete capture of audit logs of computer operating system and network packets	0.16
x_2	Maximum number of possible attacks included	0.16
x_3	Current attack behaviors	0.16
x_4	Real-world normal traffic dynamics with operation timings and industry complexity	0.16
x_5	Maintenance of cyber infrastructure performance during complete capture	0.16
x_6	Ground truth information included to assist labeling process	0.16

Table 2
Crisp input set Y.

Elements	Linguistic terms	Generation environment	$F_2(y_l)$
y_1	Good	Production or real network	1
y_2	Average	Synthetic network or testbed	0.5

(i.e., $F_1(x_k)$ and $F_2(y_l)$) can be re-enumerated after extending or defining the elements of set X and Y.

$$z_i = a[F_1(x_k)] + b[F_2(y_l)] + c \tag{1}$$

$$w_i = AndMethod\ [F_1(x_k), F_2(y_l)] \tag{2}$$

In the Sugeno fuzzy inference model, the output level of rule i (z_i) is weighted by its ring strength (w_i), as defined in Eqs. (1), (2) and (3) respectively. As the scaling output parameters (a, b and c in Eq. (1)) for the output level (z_i) are observed to be directly proportional to the number of rules (N), $a = b = c = N$. Then, the final output from the Sugeno model is the weighted average of all the rule outputs, that is,

$$Final\ \ Output = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i} \tag{3}$$

The final output is considered the initial numerical value of the realism obtained, i.e., numerical R. In order to achieve the final quantification of the R of an IDS dataset in fuzzy linguistic terms, firstly, the numerical R is normalized as $0 \le R \le 1$. Secondly, the R is considered a fuzzy linguistic variable and equals {no, low, medium, medium high, high} as a fuzzy set, with each member covering a portion of the overall values of the realism as a fuzzy linguistic term. Finally, to map the final R using fuzzy linguistic terms, the five members of fuzzy set R are related to the overlapping ranges of the probability of the realism respectively, i.e.,
$0 \le R < 0.10 \Longrightarrow$no, $0.30 \ge R > 0.08 \Longrightarrow$low,
$0.60 \ge R > 0.28 \Longrightarrow$medium,
$0.97 \ge R > 0.58 \Longrightarrow$medium high and $1 \ge R > 0.95 \Longrightarrow$high

The final values of the metric obtained from processing current datasets are shown in Table 3. Each dataset is first analyzed to observe the rules using the finding elements in the defined sets X and Y and their relationships. Based on the observed rules and their quantities N, further calculations are performed using Eqs. (1) and (2) with $i = 1 \ldots N$ to form the numerical R. The final R is obtained through normalization by dividing the numerical R by the maximum numerical R which is calculated as 12.96 using Eqs. (1), (2) and (3) respectively and realizing all the effective rules, i.e., {x_1 AND y_1, x_2 AND y_1, …, x_6 AND y_1}. In Table 3, the final R values show the probabilities of the quality of realism of existing IDS datasets. Then, to complete the fuzzification step, these values are mapped according to the fuzzy linguistic terms of fuzzy set R and plotted in Fig. 6, where each existing IDS dataset is shown linguistically as having a low or medium quality of realism.

To explain how the above metric works, we provide the following example. Let the final output from Eq. (3) be a network's performance quality, input X its management service and input Y its hardware cost. In fuzzy linguistic terms, the management service can be good, excellent or poor, the hardware cost cheap or expensive and the performance quality high, average or low. In terms of numerical values, good=8, excellent=10, poor=0, expensive=8 and cheap=2. Let us consider a rule for understanding the workings of Eqs. (1), (2) and (3) respectively, e.g., if a network's management service is poor or hardware cost cheap, its performance quality will be low. Let z and w be the numerical inference of this rule and the firing strength respectively. In order to numerically calculate z_i and w_i for the given rule using Eqs. (1) and (2), the input membership functions $F_1(x_k)$ and $F_2(y_l)$ assign the output singleton value as 0 for input X (i.e., the network's management service is poor) and 2 for input Y (i.e., the hardware is cheap). Accordingly, the final output from Eq. (3) will be 0