Environment-Related Information Security Evaluation for Intrusion Detection Systems

Ran Cheng[1(✉)], Yueming Lu[1], and Jiefu Gan[2]

[1] School of Information and Communication Engineering,
Beijing University of Posts and Communications,
Key Laboratory of Trustworthy Distributed Computing and Service (BUPT),
Ministry of Education, Beijing, China
{hscrws,ymlu}@bupt.edu.cn

[2] China Information Security Certification Center, Beijing, China
gjfsecure@163.com

Abstract. The features of actively detection of intrusion detection systems (IDSs) are crucial in cyberspace security evaluation. Most of existing evaluation models are insufficient for selecting proper IDS in varying situations since these methods only base on detection rate and false alarm ratio. The paper proposes an environment-related information security evaluation model for IDSs, and applies the model in a practical IDS evaluation process. Compared to existing ones, the proposed model considers two more factors: background traffic and workload, and thus can achieve a more objective and comprehensive evaluation result for IDSs.

Keywords: Intrusion detection system · Precision · Recall
Background traffic

1 Introduction

An increasing number of network security devices, including firewalls, intrusion detection systems (IDSs) and virtual private network (VPN) devices, etc. are implemented in widely distributed information systems nowadays. In particular, firewalls build protective barriers between internal networks and external networks; IDSs work behind the firewalls and detect potential attacks; VPNs allow encrypted information to transmit between internal networks and external networks, thus extend the range of internal networks [1]. These security devices work collaboratively for guaranteeing data to be integrate, confidential and usable.

In order to describe their relationship, ISS (Iterated Service Solution) proposed the PDR (Protection, Detection and Response) model for cyberspace security. PDR consists of protection, detection and response units. The protection unit, which is implemented by firewalls and VPNs, lags behind the

B. Li et al. (Eds.): ChinaCom 2017, LNICST 237, pp. 373–382, 2018.
https://doi.org/10.1007/978-3-319-78139-6_38

detection unit, and is not capable enough to defend against constantly changing attacks, thus causes endless system vulnerabilities. Therefore, it is not enough to keep network environments safe relying only on the protection unit. IDSs' actively detection compensates the time difference between protection and detection and builds a bridge between processes of protection and response. As the study of IDSs develops in depth, the requirements of evaluating IDSs are emerging.

Common Criteria (CC) provides a detailed set of evaluation metrics for information security products, which helps developers ensure security of their entire product development processes [2]. However, CC has two issues: 1. Evaluation results are limited to examine the documents in the scenario provided by developers; 2. evaluation time is too long.

Gan [3], and Li [4] established complete index trees for IDSs where methods including analytic hierarchy process (AHP), etc. are used to determine the weight of each index, leading to a comprehensive evaluation result. However, there are constraints among certain indices, which may lead to biased result when the result is the sum of weights.

Among those constrained indices, the performance index is the most important one when choosing an IDS since the correctness of its judgement is more measurable and more intuitive. Lincoln Laboratory of MIT conducted evaluation on performance index of IDSs back in 1998 [5]. NPV (Negative predict value), PPV (Positive predict value) and NPR (Negative Positive Ratio) metrics were extended from performance index. Those metrics are all rely solely on performance index, ignoring the different contribution of the sub-indices in performance index in the various environments of background traffic, which make them not comprehensive and objective enough under different circumstances.

Accordingly, in order to balance the constraints in the performance index, this paper proposes an environment-related evaluation model for IDSs. The model considers two supplement factors, background traffic and workload, and evaluates IDSs more comprehensively.

A comprehensive index system is still needed. According to international standard technical requirements and IDS evaluation methods, as well as the reliability requirements of IDS systems, this paper considers five quality properties including reliability, security, usability, function and performance. Upon these five properties, the paper proposes an evaluation index system for IDS devices according to their actively detection characteristics. The IDS evaluation index system is shown in Fig. 1.

Among the 5 principal indices and 12 sub-indices in Fig. 1, C_1–C_{10} are qualitative indices and the others are quantitative indices. To obtain the evaluation results, this paper uses fuzzy comprehensive evaluation (FCE) method for qualitative indices and implements the proposed model for the quantitative index, i.e., the performance index.

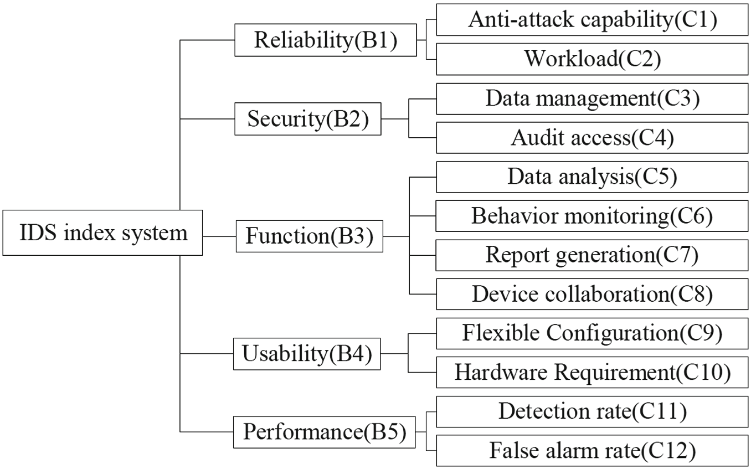

Fig. 1. IDS index system

2 Evaluation of IDS Performance

2.1 Performance Index

The most crucial factor of IDS performance evaluation is IDSs' attack detection ability and IDSs only trigger alarms when operations are judged to be abnormal. Accordingly, we can treat the IDS as a binary-classifier, which provides four possible results.

As shown in Table 1, TP (True Positive) indicates the instances where the attack is successfully detected by the system; FN (False Negative) shows the instances where the attack is not detected by the system; FP (False Positive) represents the instances where a normal operation is falsely considered as an attack by the system and TN (True Negative) is the instances where the system successfully identifies the normal operation.

Table 1. Classification results

		Predicted instances	
		Attack	Normal
Actual instances	Attack	TP	FN
	Normal	FP	TN

Accordingly, we define FAR (False alarm ratio) and DR (detection rate) as:

$$FAR = \frac{FP}{TP + FP}. \tag{1}$$

$$DR = \frac{TP}{TP + FN}. \tag{2}$$

FAR is the ratio of non-attack operations among all the operations marked abnormal by systems. DR is the fraction of intrusion operations that are detected by systems. These two indices are mainly used to evaluate IDSs.

2.2 ROC Curve

When an IDS is set to different detection thresholds, with the same input signal, the relationship between its FAR and DR can be depicted by an ROC (Receiver Operating Characteristic) curve.

In Fig. 2, curve A is a typical intrusion detection ROC curve and line B is the result of a random input. The IDS with better performance corresponds to a curve closer to the upper left corner of the graph [6].

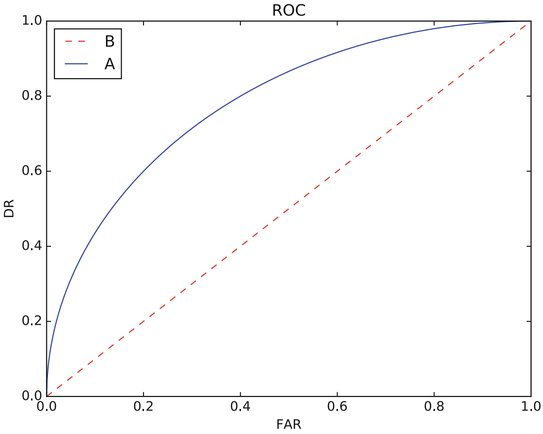

Fig. 2. Typical ROC curve

2.3 NPV and PPV Metric

This metric proposes another two metrics, i.e., NPV and PPV, which are defined from users' perspectives. PPV, also known as Bayesian detection rate, is the probability of real intrusions happening when IDSs trigger alarms. On the contrary, NPV is the probability of intrusions not happening when alarms are not triggered. However, in practice, PPV's value can be very small since the probability of the intrusion is usually low [7].

2.4 NPR Metric

NPR measures negative positive ratio [8]. The performance of an IDS can be evaluated by comparing actual NPR and predicted NPR, which is:

$$NPR = \frac{TN}{TP}. \tag{3}$$

If $NPR_P > NPR_a$, the IDS predicts more negative instances than the actual value, which means the IDS may miss attack operations and as a result DR is relatively low but FAR is high.

On the other hand, $NPR_P < NPR_a$ means FAR is relatively low but DR is high.

The NPR metric provides a new direction on evaluations of IDSs' performance, but leads to information missing at the same time.

3 An Environment-Related Performance Evaluation Model

3.1 Description of Model

In statistics, two indices, precision and recall, are used to evaluate the quality of classification results [9]. Precision is used to evaluate how accurate a system is and recall is used to evaluate how comprehensive a system is. These set of indices can be introduced into the evaluation of IDSs:

$$Precision = \frac{number\ of\ detected\ attacks}{number\ of\ alarmed\ operations} = \frac{TP}{TP+FP} = 1 - FAR. \quad (4)$$

$$Recall = \frac{number\ of\ detected\ attacks}{number\ of\ attacks\ in\ input} = \frac{TP}{TP+FN} = DR. \quad (5)$$

Precision and recall are negatively, non-linearly correlated. For example, if the system is very skeptical, and then judges most operations as attacks, its recall will be large but precision will be small. On the contrary, if the system judges few operations as attacks, its precision will be large and recall will be small. In order to integrate these two metrics into one single number, we introduce F1-score, widely used in statistical analysis, which computes the harmonic mean of precision and recall.

In practice, we consider the contribution of precision and recall in F1-score differently according to the workload of IDSs. For example, if the FAR of an IDS is too high, the credibility of the alarm decreases. Consequently, the IDS has to spend much time identifying useless information in order to get rid of false alarms, which not only costs much time but also increases workload, making system crash. Even having high DR cannot resolve this performance issue.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}. \quad (6)$$

In order for the F1-Score to be large, both precision and recall must be large meaning small FAR and large DR.

Accordingly, we give different weights to precision and recall, and the resulted performance score can be represented as follows:

$$Score_{performance} = \frac{(1+K) * (1-FAR) * DR}{1 - FAR + DR}, \quad (7)$$

and

$$K = \frac{T}{L}, \tag{8}$$

where T is the log-transformed background traffic in network environments measured in required environments, and L is the normalized score of the IDS's workload, which is computed according to the number of packages caught per second and the number of cases handled per second of the IDS.

3.2 Validation of the Model

In an environment with controlled background traffic, we measured the performances of four identical open-source IDSs. In the experiment, A and B, C and D used two different matching algorithms respectively. Their detection thresholds stay the same. We configured the method of log recording on A and C in order to get them a better recording speed and thus a higher workload score. Now given A and B's DR = 0.946, FAR = 0.182, C and D's DR = 0.86, FAR = 0.067. The results of environment-related evaluation model are as shown in Table 2.

Table 2 and Fig. 3 show the performance score of four IDSs in different background traffic environments. As shown in Fig. 3, with background traffic increasing, the scores of A and B are decreasing. This tendency does not indicate worse performance under higher background traffic, but actually means that with higher background traffic, the size of data to be processed by IDSs is larger and thus the requirement for precision is higher. As a result, the score of precision makes more contribution to the overall performance score. On the contrary, the trend for C and D increases as background traffic increases. This is because their precision score is higher than recall score.

Table 2. Implementation results

Background traffic (Mbps)	1	10	50	100	500	1000
T	0	0.3333	0.5663	0.6667	0.8997	1
AL = 0.9	0.946	0.9076	0.8920	0.8869	0.8773	0.8740
BL = 0.6	0.946	0.8959	0.8792	0.8740	0.8649	0.8617
CL = 0.9	0.863	0.8809	0.8888	0.8915	0.8966	0.8985
DL = 0.6	0.863	0.8868	0.8956	0.8985	0.9037	0.9055

We can also conclude from Fig. 3 that the slope of the curve gets lower as L increases, indicating that the IDS with higher workload depends less on precision in environments of high background traffic.

There are crossing points of four curves, which means we are not able to directly judge the performance of an IDS even with its DR and FAR given. We have to select the appropriate IDS equipment according to the environment, and this is an information that NPV, PPV, NPR evaluation metrics are not able to provide.

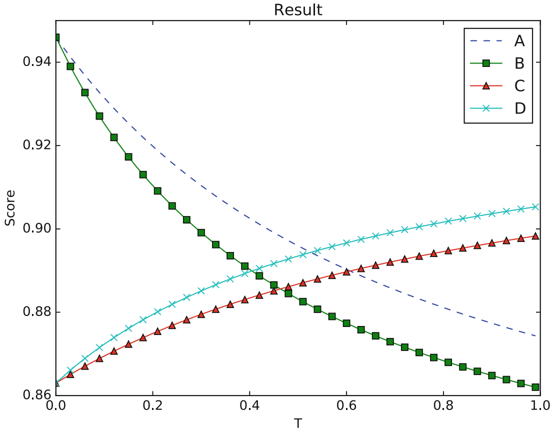

Fig. 3. Implementation results

4 Application of Evaluation Model

To give a real-world example of the effective evaluation of IDS using the environment-related performance evaluation model, we evaluate the information security level of a real-world IDS using the model with the help of the AHP and FCE.

4.1 Using the AHP to Obtain the Comparison Matrix

The AHP is about breaking a problem down and then aggregating the solution of all the sub-problems into a conclusion. A comparison matrix is built to represent the relationship between every two elements that share a common parent, and the weights of the elements will be obtained via matrix computation.

Five experts in IDS field were asked to grade the IDS's five principal indices and we use the scores to build the comparison matrix, in which each a_{ij} is an integer ranges from 1 to 9 or its reciprocal. The bigger the value, the more important index i compared to index j. The comparison matrix is showed in Table 3.

Table 3. Comparison matrix

	B1	B2	B3	B4	B5
B1	1	1/2	1/3	4	1/5
B2	2	1	1/2	5	1/4
B3	3	2	1	7	1/2
B4	1/4	1/5	1/7	1	1/8
B5	5	4	2	8	1

4.2 Matrix Consistency Test and Acquisition of Weight Vector

After finishing the comparison matrix, we test the matrix's consistency. First, we calculate the square root of the product of each row in the matrix.

$$w_1^* = \sqrt[5]{1 * \frac{1}{2} * \frac{1}{3} * 4 * \frac{1}{5}} \approx 0.668. \tag{9}$$

$$w_2^* = \sqrt[5]{2 * 1 * \frac{1}{2} * 5 * \frac{1}{4}} \approx 1.046. \tag{10}$$

$$w_3^* = \sqrt[5]{3 * 2 * 1 * 7 * \frac{1}{7}} \approx 1.838. \tag{11}$$

$$w_4^* = \sqrt[5]{\frac{1}{4} * \frac{1}{5} * \frac{1}{7} * 1 * \frac{1}{8}} \approx 0.246. \tag{12}$$

$$w_5^* = \sqrt[5]{5 * 4 * 2 * 8 * 1} \approx 3.170. \tag{13}$$

We then normalize vector w and we get $W = [0.096, 0.150, 0.264, 0.035, 0.455]$.

Next, we calculate matrix's maximum eigenvalue:

$$\lambda_{max} = \sum_{i=1}^{5} \frac{5w_i}{BW_i} = 5.196. \tag{14}$$

Thus we can obtain consistency index $CI = \frac{5.196-5}{5-1} = 0.049$. According to the table of average random index (RI) in 1–9 scale, the RI of the 5th order matrix is $RI = 1.12$ [10]. Since $\frac{CI}{RI} < 1$, the comparison matrix pass the consistency test and the weights of five principal indices are $W = [0.096, 0.150, 0.264, 0.035, 0.455]$.

Similarly, we do the AHP computation for each sub-indices and the final weight vector for C_1–C_{10} is $w = [0.04, 0.056, 0.1, 0.05, 0.039, 0.065, 0.082, 0.078, 0.021, 0.014]$.

We use the proposed model to compute B_5 as whole so its sub-indices do not need weight assignment.

4.3 FCE of Qualitative Indices

C_1–C_{10} are all qualitative indices and thus difficulty to measure accurately. To achieve the comprehensive evaluation of the system, this paper uses FCE to evaluate qualitative indices [11].

FCE is based on membership function of fuzzy mathematics, i.e., the distribution function of each index on comment set. FCE then transform the qualitative indices into quantitative form according to comment set's quantification result and thus we can achieve a comprehensive evaluation of a system restricted by various factors.

We set comment set as $V = \{excellent, good, fair, poor, very\,poor\}$ and the corresponding quantization values are $\{1, 0.75, 0.5, 0.25, 0\}$. We can obtain a fuzzy membership matrix of C_1–C_{10} using Delphi method and it is showed in (15).

$$R = \begin{bmatrix} 0.5 & 0.3 & 0.2 & 0 & 0 \\ 0.2 & 0.6 & 0.2 & 0 & 0 \\ 0.3 & 0.3 & 0.2 & 0.2 & 0 \\ 0.1 & 0.2 & 0.4 & 0.3 & 0 \\ 0 & 0.1 & 0.5 & 0.3 & 0.1 \\ 0.2 & 0.2 & 0.5 & 0.1 & 0 \\ 0 & 0 & 0.3 & 0.4 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.1 & 0 \\ 0.5 & 0.4 & 0.1 & 0 & 0 \\ 0.3 & 0.4 & 0.2 & 0.1 & 0 \end{bmatrix}. \tag{15}$$

The five elements in each row of (15) show the fuzzy distribution of each index.

C_1–C_{10}'s comprehensive evaluation vector can thus be calculated as $B = w * R = [0.1173, 0.1399, 0.1641, 0.0952, 0.0285]$. The score after Quantification is 0.328.

4.4 Calculation of Quantitative Indices' Scores

Quantitative indices are sub-indices in B5, i.e., the performance index of IDS and their scores can be calculated using (7). The results are averaged under 10 different runs of the experiment. We have $L = 0.2*1+0.6*0.75+0.2*0.5 = 0.75$, $T = 0.57$, $DR = 0.946$, $FAR = 0.182$ and the performance score is 0.78.

We combine the results of qualitative and quantitative indices and the overall score of the IDS information system is $0.328 + 0.7800.455 = 0.6829$, 'good' according to comment set.

If the background traffic in the required environment is high, in the situation where $T = 0.9$, we will have score 0.767 after recalculation, which means the tested IDS has better security detection ability in the situation where background traffic is high.

5 Conclusion and Future Work

Detection rate and false alarm ratio are both important evaluation indices to evaluate the performance of IDSs. However, in different kinds of network environments, evaluation results, obtained only upon these two indices, are not comprehensive and objective enough. By adding two supplement factors, i.e., background traffic and workload to the models utilizing DR and FAR, this paper proposed an evaluation model, which provides a new way of selecting IDSs with optimized performance in environments with different background traffic.

There is still room to improve this model. For example, there are other ways of normalizing background traffic, which makes this evaluation a qualitative but non-quantitative process. In the future, we will keep working to get a more optimized, comprehensive and objective model.

Note that it is not necessary that the proposed model be used with AHP and FCE as in this paper.

Acknowledgments. This work was supported by The Research of Key Technology and Application of Information Security Certification Project (No. 2016YFF0204001) of China Information Security Certification Center.

References

1. Stallings, W.: Network Security Essentials: Applications and Standards. Pearson Education India, Delhi (2007)
2. Herrmann, D.S.: Using the Common Criteria for IT Security Evaluation. CRC Press, Boca Raton (2002)
3. Gan, Z., He, J.: Study on multi-hierarchical fuzzy comprehensive evaluation of intrusion detection system. Appl. Res. Comput. **4**, 29 (2006)
4. Li, L., Xia, Z., Xiong, J.: Study on evaluation method of multilayer hybrid intrusion detection system. Comput. Sci. 42 (2015)
5. Lippmann, R., Fried, D., Graf, I., Haines, J., Kendall, K., McClung, D., Zissman, M.: Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation. In: Proceedings of the DARPA Information Survivability Conference and Exposition, vol. 2, pp. 12–26. IEEE (2000)
6. Haines, J., Lippmann, R., Fried, D.: 1999 DARPA intrusion detection system evaluation: design and procedures. DARPA Intrusion Detection Evaluation Design & Procedures (2001)
7. Gu, G., Fogla, P., Dagon, D., Lee, W., Skorić, B.: Measuring intrusion detection capability: an information-theoretic approach. In: Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, pp. 90–101. ACM (2006)
8. Aggarwal, P., Sharma, S.: A new metric for proficient performance evaluation of intrusion detection system. In: Herrero, Á., Baruque, B., Sedano, J., Quintián, H., Corchado, E. (eds.) International Joint Conference. AISC, vol. 369, pp. 321–331. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19713-5_28
9. Powers, D.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol. **2**, 2229–2239 (2011)
10. Saaty, L.: How to make a decision: the analytic hierarchy process. Eur. J. Oper. Res. **48**(1), 9–26 (1990)
11. Wang, X., Shi, Y., Huang, R.: Application of multi-layer fuzzy comprehensive evaluation method in debris flow assessment. J. Catastrophology **19**(2), 1–6 (2004)