IJCSMA

Gulshan Kumar, International Journal of Computer Science and Mobile Applications,

Vol.2 Issue. 11, November- 2014, pg. 11-17 ISSN: 2321-8363

Evaluation Metrics for Intrusion Detection Systems - A Study

Gulshan Kumar

Assistant Professor, Shaheed Bhagat Singh State Technical Campus, Ferozepur (Punjab)-India 152004

Email: gulshanahuja@gmail.com

Abstract

Intrusion Detection Systems (IDSs) have been evaluated using a number of ways based on different evaluation datasets for their efficiency and effectiveness. Various features of the IDSs can be evaluated, which may range from performance and correctness to usability. To evaluate different features, a large number of metrics have been proposed. Unfortunately, no benchmark metric exists till date for intrusion detection and finalizing it is still under process. Many researchers used a variety of metrics to measure the performance quantitatively.

In this paper, we explored various performance metrics used to evaluate IDSs based upon benchmark datasets. Their pros and cons are highlighted. The study in this paper will help the better understanding of different metrics for evaluating the performance of the IDSs. The findings of this paper provide useful insights into literature and are beneficial for those who are interested in applications and development of IDSs and related fields.

Keywords: Intrusions, Intrusion detection, Intrusion detection system, Network Security, Performance metrics

1. Introduction

Since the first introduction, IDSs have been evaluated using a number of ways based on different evaluation datasets (Kruegel et al., 2005). The IDS can be generally evaluated from two viewpoints (Tavallaee, 2011):

1. Efficiency: This measure deals with the resources needed to be allocated to the system including CPU cycles and main memory.
2. Effectiveness: This measure (also called classification accuracy) represents the ability of the system to distinguish between intrusive and non-intrusive activities.

Various features of the IDSs can be evaluated, which may range from performance and correctness to usability. However, most of the researchers mainly focused on measuring the accuracy and effectiveness of the IDSs in terms of false alarm rate and the percentage of attacks that are successfully detected. They did not pay much attention to the efficiency of their systems.

The validation of IDSs is generally performed by measuring benchmark metrics based on benchmark data sets followed by their comparison with other existing representative techniques in the field. Unfortunately, no benchmark metric exists till date for intrusion detection and finalizing it is still under process. As per the literature, many researchers used a variety of metrics to measure the performance quantitatively. As per statistics of a survey of 276 papers published between 2000 and 2008 conducted by (Tavallaee, 2011), 42% of the papers accessed performance of the systems by using DR, FPR and area under the ROC (AUC). But, Gu et al. highlighted that DR, FPR and AUC fails to distinguish the performance of the IDSs in some special cases (Gu et al., 2006). To overcome the limitation, they proposed a single objective metric called Intrusion Detection Capability (CID) in terms of the base rate, PPV and NPV.

Article overview: following this introduction, section 2 highlights the important performance metrics for IDS evaluation based upon benchmarked dataset proposed in literature. Description of confusion matrix and computation of metrics from confusion matrix is provided. Section 3 presents the detail of an objective metric called Intrusion Detection Capability (CID). Finally, the paper concludes the study of IDS performance metrics.

2. Performance Metrics

Several metrics have been designed to measure the effectiveness of IDS. These metrics can be divided into three classes namely threshold, ranking and probability metrics (Kumar and Kumar 2011, Caruana and Niculescu-Mizil, 2004). Threshold metrics include classification rate (CR), F-measure (FM) and Cost per example (CPE) etc. It is not important how close a prediction is to a threshold, only if it is above or below the threshold. The value of threshold metrics lies in the range from 0 to 1. Ranking metrics include False Positive Rate (FPR), Detection Rate (DR), Precision (PR), Area under ROC curve (AUC) and Intrusion Detection Capability (CID). The value of ranking metrics lies in the range from 0 to 1. These metrics depend on the ordering of the cases, not the actual predicted values. As long as the ordering is preserved, it makes no difference. These metrics measure how well the attack instances are ordered before normal instances and can be viewed as a summary of model performance across all possible thresholds. Probability metrics include root mean square error (RMSE). The value of RMSE lies in the range from 0 to 1. The metric is minimized when the predicted value for each attack class coincides with the true conditional probability of that class being normal class. However, CID is an information theory based metric which gives a better comparison of various IDSs than the other popular metric like AUC (Gu et al. 2006). The value of CID lies between 0 and 1. Higher the value of the CID better is the performance of the IDS. Generally, these metrics are computed from confusion matrix. The confusion matrix is the best way to represent classification results of the IDS

2.1 Confusion matrix

Confusion matrix is a matrix that represents result of classification. It represents true and false classification results. The followings are the possibilities to classify events and depicted in Table 1:

– True positive (TP): Intrusions that are successfully detected by the IDS.

– False positive (FP): Normal/non-intrusive behavior that is wrongly classified as intrusive by the IDS.

– True Negative (TN): Normal/non-intrusive behavior that is successfully labeled as normal/non-intrusive by the IDS.

– False Negative (FN): Intrusions that are missed by the IDS, and classified as normal/non-intrusive.

Actual	Predicted	Predicted
	Attack	Normal
Attack	TP	FN
Normal	FP	TN

Table 1: Confusion matrix

2.2 Metrics from confusion matrix

In spite of the representational power of the confusion matrix in classification, it is not a very useful tool for the sake of comparison of the IDSs. To solve this problem, different performance metrics are defined in terms of the confusion matrix variables. These metrics produce some numeric values that are easily comparable and are briefly explained in subsequent paragraphs.

1. Classification rate (CR): It is defined as the ratio of correctly classified instances and the total number of instances.

$$CR = \frac{Correctlyclassifiedi ns \tan ces}{Tota \ln umberofins \tan ce}$$

$$= \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

2. Detection rate (DR): It is computed as the ratio between the number of correctly detected attacks and the total number of attacks.

$$DR = \frac{Correctly \det ectedattak s}{Tota \ln umberofattacks}$$

$$= \frac{TP}{TP+FN} \qquad (2)$$

3. False positive rate (FPR): It is defined as the ratio between the number of normal instances detected as attack and the total number of normal instances.

$$FPR = \frac{Numberofnormalins \tan ces \det ectedasatt acks}{Tota \ln umberofnormalins \tan ces}$$

$$= \frac{FP}{FP+TN} \qquad (3)$$

4. Precision (PR): It is the fraction of data instances predicted as positive that are actually positive.

$$PR = \frac{TP}{TP+FP} \qquad (4)$$

5. Recall: This metric measures the missing part from the Precision; namely, the percentage from the real attack instances covered by the classifier. Consequently, it is desired for a classifier to have a high recall value. This metric is equivalent to the detection rate (DR).
6. F-measure (FM): For a given threshold, the FM is the harmonic mean of the precision and recall at that threshold.

$$FM = \frac{2}{\frac{1}{PR}+\frac{1}{\mathrm{Re}\,call}} \qquad (5)$$

F-Measure is preferred when only one accuracy metric is desired as an evaluation criterion.

7. Area under ROC curve (ROC): ROC is a plot of sensitivity vs. (1-specificity) for all possible thresholds. Sensitivity is defined as P (Pred = positive |True = positive) and is approximated by the fraction of true positives that are predicted as positive (this is the same as a recall). Specificity is P (Pred = negative |True = negative). It is approximated by the fraction of true negatives predicted as negatives. Area under the ROC curve is used as a summary statistic. Originated from signal detection theory (Tavallaee, 2011), ROC curves are used on the one hand to visualize the relation between detection rate and false positive rate of a classifier while tuning it, and on the other hand to compare the accuracy of several classifiers. Although this measure is very effective, it has some limitations. The first limitation is that it is dependent on the ratio of attacks to normal traffic. The

comparison of various classifiers based upon ROC works fine for the same dataset. However, the comparison of the IDSs done on various data sets is completely wrong, unless they have the same ratio of attack to normal instances. The second problem with ROC curves is that they might be misleading and simply incomplete for understanding the strengths and weaknesses of the candidate system (McHugh, 2000, Axelsson, 2000).

Thus, in order to evaluate the effectiveness of an IDS, we need to measure its ability to correctly classify events as normal or intrusive along with other performance objectives, such as the economy in resource usage, resilience to stress and ability to resist attacks directed at the IDS (Gu et al., 2006). Measuring these abilities of IDS is important to both industry as well as the research community. It helps us to tune the IDS in a better way as well as compare different IDSs. As discussed above, there exist many metrics that measure different aspects of IDS, but no single metric seems sufficient to measure the capability of the IDSs objectively. As per statistics of a survey conducted by (Tavallaee, 2011), the most widely used metrics by the intrusion detection research community are True Positive Rate (TPR) and False Positive Rate (FPR) along with the ROC. False Negative rate FNR = 1-TPR and True Negative Rate TNR = 1-FPR can also be used as an alternate. Based upon values of these two metrics only, it is very difficult to determine better IDS among different IDSs especially when the tradeoff is needed. For example, one IDS is reporting, TPR = 0.8, FPR = 0.1, while at another IDS, TPR = 0.9, FPR =0.2. If only the metrics of TPR, FPR are given, it is very difficult to determine the better IDS. No doubt, the ROC curve provides tradeoff, but it cannot tell which one is better in many cases (Gu et al., 2006). To solve this problem, (Gu et al., 2006) proposed a new objective metric based upon information theory called Intrusion Detection Capability (CID) considering the base rate, TPR and FPR collectively. CID possesses many important features. For example, 1) it naturally takes into account all the important aspects of detection capability, i.e., FPR, FNR, positive predictive value (PPV), negative predictive value (NPV), and base rate (the probability of intrusions); (2) it objectively provides an essential measure of intrusion detection capability; (3) it is very sensitive to the IDS operation parameters such as base rate, FPR and FNR.

3. Intrusion Detection Capability (CID) metric

Sometimes it is difficult to determine which IDS is better than another in terms of only FPR and TPR. For example, IDS1 can detect 10% more attacks, but IDS2 can produce 10% lower false alarms. Which one is better? In order to solve the problem, (Gu et al., 2006) suggested a single unified objective metric called intrusion detection capability (CID) based upon base rate, positive predictive value, or Bayesian detection rate (PPV) and negative predictive value (NPV). Such metric is used to select the best IDS configuration for an operational environment and to evaluate different IDSs.

$$CID = \frac{I(X;Y)}{H(X)}$$

$$= \frac{H(X) - H(X\,|\,Y)}{H(X)} \qquad (6)$$

Where I (X; Y) give the mutual information of X and Y, H (X) gives the entropy of X and H (X |Y) gives the conditional entropy of X after Y is known.

$$H(X) = -\sum_{x} p(x)\log(p(x)) = -B\log(B) - (1-B)(\log(1-B)) \qquad (7)$$

In terms of base rate (B), PPV and NPV, the CID can be computed as

$$CID = -B(1-\beta)\log(PPV) - B(1-\beta)\log(1-NPV) - (1-B)(1-\alpha)\log(NPV) - (1-B)\alpha\log(1-PPV) \qquad (8)$$

Detail of CID can be further studied in (Gu et al., 2006).

4. Conclusions

The aim of this paper is to present various performance metrics used for evaluation of an IDSs based upon benchmark datasets and their limitations. The paper introduced need and significance of performance metrics. Various metrics are explored to measure the performance of IDSs and validate them. The validation of IDSs is generally performed by measuring benchmark metrics based on benchmark data sets followed by their comparison with other existing representative techniques in the field. Unfortunately, no benchmark metric exists till date for intrusion detection and finalizing it is still under process. Many researchers used a variety of metrics to measure the performance quantitatively. As per literature, most of the papers accessed performance of the IDSs by using DR, FPR and area under the ROC (AUC). But, it is highlighted that DR, FPR and AUC fails to distinguish the performance of the IDSs in some special cases. To overcome the limitation, a single objective metric called Intrusion Detection Capability (CID)is proposed in terms of the base rate, PPV and NPV. Keeping these points into consideration, we suggest to used DR, FPR and CID along with DR of each attack class to compare the performance of the IDSs with the existing techniques.

References

1. Axelsson, S.: Intrusion detection systems: A survey and taxonomy. Tech. rep., Technical report (2000)

2. Caruana, R., Niculescu-Mizil, A.: Data mining in metric space: an empirical analysis of supervised learning performance criteria. In: Proc. of the tenth ACM SIGKDD international conference on Knowledge

discovery and data mining, pp. 69–78. ACM (2004)

3. Gu, G., Fogla, P., Dagon, D., Lee, W., Skorić, B.: Measuring intrusion detection capability: An information-theoretic approach. In: Proc. of the 2006 ACM Symposium on Information, computer and communications security, pp. 90–101. ACM (2006)

4. Kruegel, C., Vigna, G., Robertson, W.: A multi-model approach to the detection of web- based attacks. Computer Networks 48(5), 717–738 (2005)

5. Kumar, G., Kumar, K.: Ai based supervised classifiers: an analysis for intrusion detection.

In: Proc. of International Conference on Advances in Computing and Artificial Intelligence, pp. 170–174. ACM (2011)

6. McHugh, J.: Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by lincoln laboratory. ACM transactions on Information and system Security 3(4), 262–294, (2000)

7. Tavallaee, M.: An adaptive hybrid intrusion detection system. Ph.D. thesis, University of new brunswick (2011)

Author Biography

Dr. Gulshan Kumar has received his MCA degree from Guru Nanak Dev University Amritsar (Punjab) India in 2001, and M.Tech. Degree in computer science & engineering from JRN Rajasthan Vidyapeeth Deemed University, Udaipur (Rajasthan)-India, in 2009. He got his Ph.D. from Punjab Technical University, Jalandhar (Punjab)-India. He has 12 year of teaching experience. He has 30 international and national publications to his name. Currently, he is working as Assistant Professor in Computer Applications department at Shaheed Bhagat Singh State Technical Campus, Ferozepur (Punjab)-India. He has supervised 03 M. Tech. students for their final thesis. His current research interests involve Artificial Intelligence, Network Security, Machine Learning and Databases.