ROC in Assessing IDS Quality

Rune Hammersland
{firstname.lastname}@hig.no
Norwegian Information Security Lab, Gjøvik University College

November 30, 2007

1 Terms

For assessing the quality of IDS systems, we first need to define a couple of terms. When an IDS is looking at the network traffic (or events on the host for host based IDSes), it tries to decide if the traffic is is malicious or not. When the IDS indicates an intrusion, this is called a Positive. When the IDS indicates an intrusion and an intrusion indeed is in progress, we have a True Positive (TP). In other words the positive assertion from the IDS is to be trusted, as it is a correct assertion. When the IDS indicates an intrusion even though no attack is in progress, we have a False Positive (FP).

On the other hand, when the IDS indicates that the traffic or event is harmless, this is called a Negative. When the IDS indicates that no intrusion is in progress and this is a correct assertion, we have a True Negative (TN). And the opposite: when the IDS indicates that no intrusion is in progress, while there actually is an intrusion attempt, we have a False Negative (FN). These terms are shown in Table 1. From here on out, we shall assume a Network Intrusion Detection System.

The total number of intrusions during a certain amount of time is defined as the number of True Positives and the number of False Negatives (as these are actually intrusions). The total number of non-intrusions is thus derived from the number of True Negatives and the number of False Positives (as these actually aren't intrusions). From these numbers we can

Table 1: Truthtable for intrusion assertion by an IDS

		Assertion by the IDS	
		Positive	Negative
Intrusion?	Yes	True Positive	False Negative
	No	False Positive	True Negative

then derive the Base Rate, which is the probability of an attack on our system. This is defined as the number of intrusions over the number of events. See also Equation 1

$$BaseRate = \frac{TP + FN}{TP + FP + TN + FN} \tag{1}$$

Other rates that are interesting when evaluating an IDS are True Positive Rate, False Positive Rate, True Negative Rate and False Negative Rate. These rates corresponds to the terms already defined (the same names, only without the word Rate). While a True Positive is an actual intrusion correctly detected by the IDS, the True Positive Rate is the probability that an intrusion attempt in the environment is detected by the IDS (see i.e. [Pet06]). In other words, the number of True Positives over all intrusion attempts in the environment:

$$TPR = P(A|I) = \frac{TP}{TP + FN} \tag{2}$$

where A is an Alarm and I is an Intrusion. The False Positive Rate, the probability of an alarm given no intrusion, is given by:

$$FPR = P(A|\neg I) = \frac{FP}{FP + TN} \tag{3}$$

The probability of no alarm given no intrusions, the True Negative Rate is given by: $TNR = P(\neg A|\neg I) = \frac{TN}{FP+TN}$. Lastly, the False Negative Rate, the probability of no alarm given an actual intrusion, is given by: $FNR = P(\neg A|I) = \frac{FN}{TP+FN}$.

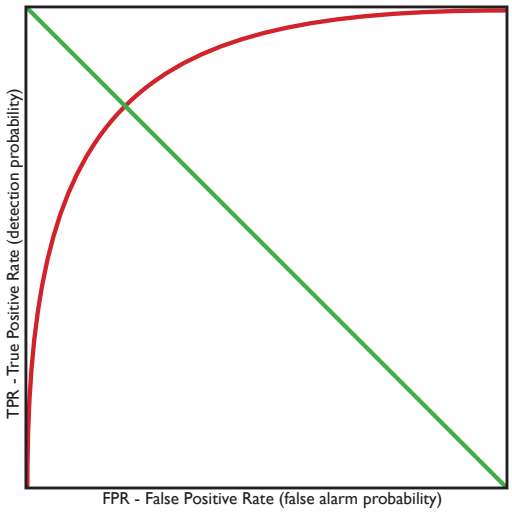

Figure 1: Example of a ROC curve. The red line represents the trade-off between TPR and FPR. The intersection between the red and the green line is where we find the Equal Error Rate.

2 ROC - Receiver Operating Characteristics

ROC curves are often used to identify the quality of an automated system where some criteria can be changed in order to affect the quality of the system. The curve is a plot between hit rates and false alarm rates in order to show the trade-off between them. Systems for biometric authentication for instance often use ROC curves to find the Equal Error Rate — the point where the usability of the system is as good as the security. This point is identified by plotting the False Match Rate (the number of impostors entering as a legitimate user) against the False Non-Match Rate (the number of legitimate users turned down by the system) as a ROC curve (indicated by a red line in Figure 1). Using that curve, we find the Equal Error Rate by looking at the intersection between the curve and the linear function $y = y_{max} - x$ (indicated by a green line in Figure 1).

ROC curves sees a great amount of use in digital signal processing, medical decision making, evaluation of radiologic systems and machine learning (see [Faw03], which also includes a discussion of misconceptions and pitfalls regarding to the use of ROC curves). Apparently, Lippmann et al. [LFG+00], in their extensive 1998 (and 1999) study, were the first ones

to apply ROC curves to evaluate different kinds of IDS.

In assessing an IDS, we are usually interested in looking at how changes in the ruleset affect the detection probability (TPR) and the probability for a false alarm (FPR) [MHL+03]. That way we can find out how to define the ruleset in a way that yields high detection probability, while maintaining the false alarm probability as low as reasonably possible. A problem with this approach, is that it doesn't take into account that the False Negatives usually have higher cost than False Positives [OCR06].

Gaffney and Ulvila [GU01] has proposed a method for evaluating an IDS which is based on a combination of a ROC analysis and a cost analysis. It also tries to assess the hostility of the operating environment. The method implements this through a decision tree, and calculates expected costs. After doing this, it tries to find an optimal operating point. As they take into account the cost of false alarms and detection failures as well as the probability of an attack and the ROC curve, they argue that their approach better describes the quality of an IDS.

Axelsson [Axe00] states that in creating a ROC curve for an IDS, we will always touch upon the points $(0, 0)$ and $(1, 1)$, as we can always decide that all traffic is either malicious or benign. Further he states that if we make a random decision about whether the traffic is malicious or not, the system will achieve TPR and FPR somewhere on the diagonal between $(0, 0)$. Due to this fact we can say that any sensible IDS should perform above this diagonal.

3 Usefullness of ROC Curves for IDS

In general it is very hard to test an IDS, because you can only simulate traffic and attacks, or put the system in the production environment to test it with real data. In the first case you can only test the system against known attack patterns, and these are patterns you most likely have written good rules for anyway. The only benefit you get from those tests is that you can make sure those rules indeed are effective, and that the IDS doesn't take too long time to process the data (which could lead to packet drops, which again leads to higher FNR). In the latter case you would test the system against real data, however it is much harder to establish what traffic is really part of an attack. Mell et al. [MHL+03] points to weak signatures for signature based IDS that i.e. triggers on all traffic to a certain port, detec-

tion of common violations of the TCP protocol, and of course that there is no such thing as a "standard" network. All of these things makes it hard to test an IDS, and to verify whether you have True or False Positives when the IDS issues an alarm, as well as the opposite aspect: identifying the False Negative Rate.

Furthermore, McHugh in his critique [McH00] on IDS evaluations performed by Lippmann et al., claims that ROC curves have "unanticipated problems in it's application to IDS evaluation". He lists the following problems: "[...] problems in determining appropriate units of analysis, bias towards possibly unrealistic detection approaches, and questionable presentations of false alarm data." We can imagine the temptation to tune the rule set to better match the testing environment, which might not be the same as the production environment.

To be able to use ROC curves to compare two or more systems, the curves should be generated using similar, or preferrably the same test data. McHugh points to problems that arises when systems are tested against predefined corpora: "[the corpus requires] careful construction and validation if their content and structure are not to bias the systems that are developed using them for test data" (in other words: systems tuned to perform well using the test data).

When comparing ROC curves for two Intrusion Detection Systems, Orfila et al. [OCR06] states that as long as the two curves intersects at some point, we cannot claim that one system is better than the other. The reason for this is that while some users will need a high detection rate (and therefore are partial to the system which performs on the top right part of the ROC curve) another user might need a low false alarm probability rate (the system that performs on the bottom left part of the curve). The needs of users can of course be placed anywhere on the top left part of the diagonal, so it is highly subjective which system is "best". When we have two systems, A and B, and the following condition is met:

$$\forall(x_a, y_a), (x_b, y_b) \in A, B : y_a >= y_b \tag{4}$$

where (x_a, y_a) is a point on the curve of system A, x_a is a point on the x-axis (FPR) and y_a a point on the y-axis (TPR), we can conclude that system A outperforms system B, as A will be as good, or better for every test. A user can then safely choose system A, regardless of need.

As we can see, the ROC curve only applies to how the system per-

formed in the testing environment. In this enviromnent we test the system for it's detection rate for known attacks, and while this is valuable, there is no way to test how well the system handles new attacks (unless the system, and rule set was made before a new attack method). This method of testing might give a bias towards signature based systems, as these generally contains signatures for all known attacks, while an anomaly based system probably will perform better. Axelsson also points to this in his conclusion [Axe00], stating that anomaly based systems often will yield higher false alarm rates, while the cost of a false alarm usually is lower than the cost of false negatives.

References

[Axe00] Stefan Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, 3(3):186–205, 2000.

[Faw03] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers, 2003.

[GU01] Jr. Gaffney, J.E. and J.W. Ulvila. Evaluation of intrusion detectors: a decision theory approach. *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on*, pages 50–61, 2001.

[LFG+00] R. Lippmann, D. Fried, I. Graf, J. Haines, K. Kendall, D. McClung, D. Weber, S. Webster, D. Wyschogrod, R. Cunningham, and M. Zissman. Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation. *DARPA Information Survivability Conference and Exposition, 2000. DISCEX '00. Proceedings*, 2:12–26, 2000.

[McH00] John McHugh. 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.

[MHL+03] P. Mell, V. Hu, R. Lipmann, J. Haines, and M. Zissman. An overview of issues in testing intrusion detection systems, June 2003.

[OCR06] Agustín Orfila, Javier Carbó, and Arturo Ribagorda. *Advances in Data Mining*, volume 4065, chapter Effectiveness Evaluation of Data Mining Based IDS, pages 377–388. Springer Berlin / Heidelberg, 2006.

[Pet06] Slobodan Petrovic. A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In *Proceedings of the 11th Nordic Workshop on Secure IT-systems, NORDSEC 2006*, pages 53–64, 2006.