An Information-Theoretic Measure of Intrusion Detection Capability

Guofei Gu, Prahlad Fogla, David Dagon, Wenke Lee
College of Computing, Georgia Institute of Technology, Atlanta GA 30332
{guofei,prahlad,dagon,wenke}@cc.gatech.edu
Boris Škorić
Philips Research Laboratories, Netherlands
boris.skoric@philips.com

Abstract

A fundamental problem in intrusion detection is what metric(s) can be used to objectively evaluate an intrusion detection system (IDS) in terms of its ability to correctly classify events as normal or intrusion. In this paper, we provide an in-depth analysis of existing metrics. We argue that the lack of a single unified metric makes it difficult to fine tune and evaluate an IDS. The intrusion detection process can be examined from an information-theoretic point of view. Intuitively, we should have less uncertainty about the input (event data) given the IDS output (alarm data). We thus propose a new metric called *Intrusion Detection Capability*, C_{ID}, which is simply the ratio of the mutual information between IDS input and output, and the entropy of the input. C_{ID} has the desired property that: (1) it takes into account all the important aspects of detection capability naturally, i.e., true positive rate, false positive rate, positive predictive value, negative predictive value, and base rate; (2) it objectively provide an intrinsic measure of intrusion detection capability; (3) it is sensitive to IDS operation parameters. We propose that C_{ID} is the appropriate performance measure to maximize when fine tuning an IDS. The thus obtained operation point is the best that can be achieved by the IDS in terms of its intrinsic ability to classify input data. We use numerical examples as well as experiments of actual IDSs on various datasets to show that using C_{ID}, we can choose the best (optimal) operating point for an IDS, and can objectively compare different IDSs.

1 Introduction

A fundamental problem in intrusion detection is what metric(s) can be used to objectively measure the effectiveness of an intrusion detection system (IDS) in terms of its ability to correctly classify events as normal or intrusion. Defining an appropriate metric is essential to both practice and research because we need a metric when selecting the best IDS configuration for an operation environment and when comparing different IDSs.

The most basic and commonly used metrics are true positive rate (TP), which is the probability that the IDS outputs an alarm when there is an intrusion, and false positive rate (FP), which is the probability that the IDS outputs an alarm when there is no intrusion. Alternatively, one can use false negative rate $FN = 1 - TP$ and true negative rate $TN = 1 - FP$. When we fine tune an IDS (especially an anomaly detection system), for example by setting the threshold of deviation from a normal profile, there may be different TP and FP values associated with different IDS operation points (e.g., each with a different threshold). Clearly, both TP and FP need to be considered when selecting the best IDS operation point and comparing IDSs. The question is then how to use these two metrics together.

A popular approach is to use an ROC (receiver operating characteristic) curve [HW66] to plot the different TP and FP values associated with different IDS operation points. For example, an ROC curve can show one (operation) point with $<TP = 0.99, FP = 0.001>$ and another with $<TP = 0.999, FP = 0.01>$, etc. An ROC curve shows the relationship between TP and FP but by itself cannot be used to determine the best IDS operation point. ROC curves may be used for comparing IDSs. If the ROC curves of two IDSs do not "cross" (i.e., one is *always* above the other), then the IDS with the top ROC curve is better because for every FP it has a higher TP. However, if the curves do cross, then there is no easy way to compare the IDSs. It is not

Term	**Equivalent Terms from IDS Literature**	**Meaning**
FP, or α	$P(A\|\neg I)$	False positive rate. The chance that there is an alert, A, when there is no intrusion, $\neg I$.
TP	$(1-\beta)$, $P(A\|I)$	True positive rate. The chance the there is an alert, A, when there is an intrusion, I.
FN, or β	$P(\neg A\|I)$	False negative rate. The chance there is no alert, $\neg A$, when there *is* an intrusion, I.
TN	$(1-\alpha)$, $P(\neg A\|\neg I)$	True negative rate. The chance there is no alert, $\neg A$, when there is *no* intrusion, $\neg I$.
PPV	"Bayesian detection rate", $P(I\|A)$	Positive predictive value. The chance that an intrusion, I, is present when an IDS outputs an alarm, A.
NPV	$P(\neg I\|\neg A)$	Negative predictive value. The chance that there is no intrusion, $\neg I$, when an IDS does not output an alarm, $\neg A$.
B	$P(I)$	Base rate. The probability that there is an intrusion in the observed audit data.

Table 1: List of terminology used in this paper. For readability, we will use the terms listed in the leftmost column.

always appropriate to use the area under ROC curve (AUC) for comparison because it measures all possible operation points of an IDS. One can argue that comparison should be based on the best operation point of each IDS because in practice an IDS is fine tuned to a particular configuration (e.g., using a particular threshold).

One approach to integrate the metrics TP and FP together is through cost-based analysis. Essentially, the tradeoff between a true positive and a false positive is considered in terms of cost measures (or estimates) of the damage caused by an intrusion and inconvenience caused by a false alarm. Gaffney and Ulvila [GU01, UG03] use such an approach to combine ROC curves with cost analysis to compute an expected cost for each IDS operation point. The expected cost can be used to select the best operation point and to compare different IDSs. The quality of cost-based analysis depends on how well the cost estimates reflect the reality. However, cost measures in security are often determined *subjectively* because of the lack of good (risk) analysis models. Thus, cost-based analysis cannot be used to *objectively* evaluate and compare IDSs. Moreover, cost-based analysis does not give an intrinsic measure of detection performance (or accuracy).

In addition to TP and FP, two other useful metrics are the positive predictive value (PPV), which is the probability that there is an intrusion when the IDS outputs an alarm, and negative predictive value (NPV), which is the probability that there is no intrusion when the IDS does not output an alarm. These metrics are very important from a usability point of view because ultimately, the IDS alarms are useful to an intrusion response system (or admin staff) only if the IDS has high PPV and NPV. Both PPV and NPV depend on TP and FP, and are very sensitive to base rate (B), which is the prior probability of intrusion. Thus, these two metrics can be expressed using Bayes theorem (and PPV is called Bayesian detection rate [Axe99] in IDS literature) so that the base rate can be entered as a piece of prior information about the IDS operational environment in the Bayesian equations. Similar to the situation with TP and FP, both PPV and NPV are needed when evaluating an IDS from a usability point of view, and currently there is no objective method to integrate both metrics.

We need a single unified metric that takes into account all the important aspects of detection capability, i.e., TP, FP, PPV, NPV, and B. That is, this metric should incorporate existing metrics because they all are useful in their own rights. This metric needs to be objective. That is, it should not depend on any subjective measure. In addition, it needs to be sensitive to IDS *operation parameters* to facilitate fine tuning and fine-grained comparison of IDSs. We use TP and FP as the surrogates of IDS operation parameters (e.g., threshold) because changing the operation parameters usually results in changes to TP and FP. Although it is difficult or sometimes impossible to control the base rate in an IDS, we still consider it as an operation parameter because it is a measure of the environment in which the IDS operates. TP, FP, B can be measured when we evaluate an IDS because we have the evaluation data set and should know the ground truth.

We propose an information-theoretic measure of intrusion detection capability. At an abstract level, the purpose of an IDS is to classify the input data (i.e., events that the IDS monitors) correctly as normal or an intrusion. That is, the IDS output (i.e., the alarms) should faithfully reflect the "truth" about the input (or whether there is an intrusion or not). From an information-theoretic point of view, we should have less *uncertainty* about the input given the IDS output. Thus, our metric, called *Intrusion Detection Capability*, or C_{ID}, is simply the ratio of the mutual information between IDS input and output, and the entropy of the input. Mutual information measures the amount of uncertainty of the input resolved by knowing the IDS output. We normalize it using the entropy (the original uncertainty) of the input. Thus, the ratio provides a normalized measure of the amount of certainty gained by observing IDS outputs. This natural metric incorporates TP, FP, PPV, NPV and B, and thus provides a unified measure of the detection capability of an IDS. It is also sensitive to TP, FP, and B.

This paper makes contributions to both research and practice. We provide an in-depth analysis of existing metrics and provide a better understanding of their limitations. We examine the intrusion detection process from an information-theoretic point of view and propose a new unified metric for intrusion detection capability. C_{ID} is the appropriate performance measure to maximize when fine tuning an IDS. The thus obtained operation point is the best that can be achieved by the IDS in terms of its intrinsic ability to classify input data. We use numerical examples as well as experiments of actual IDSs on various datasets to show that using this metric, we can choose the best (optimal) operating point for an IDS, and can objectively compare different IDSs.

Note that this new metric, C_{ID}, is not intended to replace existing metrics such as TP, FP, etc. In fact, TP, FP are used as basic inputs to compute C_{ID}. Thus, C_{ID} presents a composite/unified measure. Furthermore, C_{ID} is just one possible measure for IDS evaluation. Other approaches, e.g., cost-based analysis, may be appropriate in some cases, e.g., when good (risk) analysis models are available.

Also note that in this paper we are not concerned with other important IDS performance objectives, such as economy in resource usage, resilience to stress [PZC$^+$96], and ability to resist attacks directed at the IDS [PN98, Pax99]. We only focus on the intrinsic IDS ability of classifying input events accurately. Although our measure can be used in other domains, we focus on intrusion detection (specifically network-based intrusion detection) as a motivating example.

The rest of the paper is organized as follows. In section 2, we provide an information-theoretic view of the intrusion detection process. After reviewing some essential information theory concepts, we introduce our unified metric of intrusion detection capability, C_{ID}. In section 3, we analyze existing metrics and compare them with C_{ID}. Section 4 describes how C_{ID} can be used to select the best operation point of an IDS and to compare different IDSs. Section 5 discusses limitations and extensions. Section 6 introduces related work. We conclude the paper and discuss future work in Section 7.

2 An Information-Theoretic View of Intrusion Detection

Let us revisit the intrusion detection process from an information-theoretic point of view. At an abstract level, an IDS accepts and analyzes an input data stream, and produces alerts to indicate intrusions. Every unit of input data stream has the status of either intrusion or normal. Thus, we can model the input of an IDS as a random variable X, where $X = 1$ represents an intrusion, and $X = 0$ represents normal traffic. The output alerts of an IDS is also modeled as a random variable Y, where $Y = 1$ means there is an alert indicating an intrusion, and $Y = 0$ represents no alert from the IDS. We assume here that there is an IDS output (decision) corresponding to each input. The exact encoding of X, Y is related to the unit of input data stream, which is in fact related to IDS analysis granularity, or the so-called unit of analysis [McH00]. For network-based IDSs such as Snort [Roe99], the unit of analysis is a packet. The malicious packets are encoded as $X = 1$. The IDS examines every packet to classify it as malicious ($Y = 1$) or normal ($Y = 0$). There are also IDSs such as Bro [Pax99] which analyze events based on flows. In this case, the malicious flow is encoded as $X = 1$ and the output indicates whether this flow contains an attack ($Y = 1$) or not ($Y = 0$).

An abstract model for intrusion detection is shown in Figure 1. In this model, $p(X = 1)$ is the base rate,

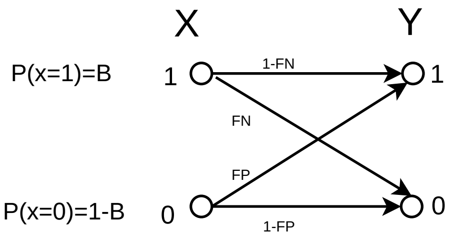


Figure 1: An abstract model for intrusion detection.

which is the prior probability of intrusion in the input event data examined by the IDS. We denote it as B. An intrusion event has a probability $p(Y = 0|X = 1)$ of being considered normal by the IDS. This is the false negative rate (FN) and is denoted as β. Similarly, a normal event also has a probability $p(Y = 1|X = 0)$ of being misclassified as an intrusion. This is the false positive rate (FP) and is denoted as α. We will use the notations (B, α, β) throughout this paper. Table 1 lists the terms used by this paper and their meaning. Note that when we evaluate an IDS, we should have the evaluation data set of which we know the ground truth. So once the evaluation data set is given and the tests are run, we should be able to calculate B, α and β.

This model is very useful because it allows us to analyze intrusion detection from an information-theoretic point of view. We will first review a few basic concepts in information theory [CT91], which are the building blocks of our proposed metric of intrusion detection capability.

2.1 Information Theory Background

Definition 1 *The entropy (or self-information) H(X) of a discrete random variable X is defined by [CT91]*

$$H(X) = -\sum_x p(x) \log p(x)$$

This definition is commonly known as the Shannon entropy measure, or the uncertainty of X. A larger value of $H(X)$ indicates that X is more uncertain. We use the convention that $0 \log 0 = 0$, which is easily justified by continuity because $x \log x \to 0$ as $x \to 0$. The definition of entropy can be extended to the case of jointly distributed random variables.

Definition 2 *If (X, Y) is jointly distributed as $p(x, y)$, then the joint entropy H(X,Y) of X and Y is defined by [CT91]*

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log p(x, y)$$

Definition 3 *If (X, Y) is jointly distributed as $p(x, y)$, then the conditional entropy $H(X|Y)$ is defined as [CT91]*

$$\begin{aligned} H(X|Y) &= \sum_y p(y) H(X|Y = y) = -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \\ &= -\sum_y \sum_x p(x, y) \log p(x|y) = -\sum_x \sum_y p(x)p(y|x) \log \frac{p(x,y)}{p(y)} \end{aligned} \tag{1}$$

Conditional entropy is the amount of remaining uncertainty of X after Y is known. We can say $H(X|Y) = 0$ if and only if the value of X is completely determined by the value of Y. Conversely, $H(X|Y) = H(X)$ if and only if X and Y are completely independent. Conditional entropy $H(X|Y)$ has the following property:

$$0 \leq H(X|Y) \leq H(X)$$

Definition 4 *Consider two random variables X and Y with a joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is defined as [CT91]*

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual information tells us the amount of information shared between two random variables X and Y. Obviously, $I(X;Y) = I(Y;X)$.

Theorem 1 *Mutual information and entropy [CT91]:*

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

This shows that we can interpret mutual information as the amount of *reduction* of uncertainty in X after Y is known, $H(X|Y)$ being the *remaining* uncertainty. This theorem shows the relationship between conditional entropy and mutual information. We can also express this relationship in a Venn diagram as shown in Figure 2. Here, mutual information $I(X;Y)$ corresponds to the intersection of the information in X with the information in Y. Clearly, $0 \leq I(X;Y) \leq H(X)$.

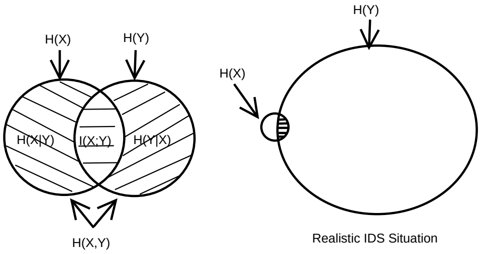


Figure 2: Relationship between entropy and mutual information. For example, the entropy of X, or $H(X)$, is the sum of the mutual information of X and Y, or $I(X;Y)$, and the conditional entropy between the two, or $H(X|Y)$. On the right, the entropy $H(Y)$ is much larger than $H(X)$. This reflects a likely IDS scenario, where the base rate is very small (close to zero), so $H(X)$ is nearly zero. On the other hand, the IDS may produce quite a few false positives. Thus, $H(Y)$ can be larger than $H(X)$.

2.2 C_{ID}: A New Metric of Intrusion Detection Capability

Our goal is to define a metric to measure the capability of an IDS to classify the input events correctly. At the abstract level, the purpose of an IDS is to classify the input correctly as normal or intrusion. That is, the IDS output should faithfully reflect the “truth” about the input (or whether there is an intrusion or not). From an information-theoretic point of view, we should have less *uncertainty* about the input given the IDS output. Mutual information is a proper yardstick because it captures the reduction of original uncertainty (intrusion or normal) given that we observe the IDS alerts.

We propose a new metric, *Intrusion Detection Capability*, or C_{ID}, which is simply the ratio of the mutual information between IDS input and output, and the entropy of the input.

Definition 5 *Let X be the random variable representing the IDS input and Y be the random variable representing the IDS output. Intrusion Detection Capability is defined as*

$$C_{ID} = \frac{I(X;Y)}{H(X)} \tag{2}$$

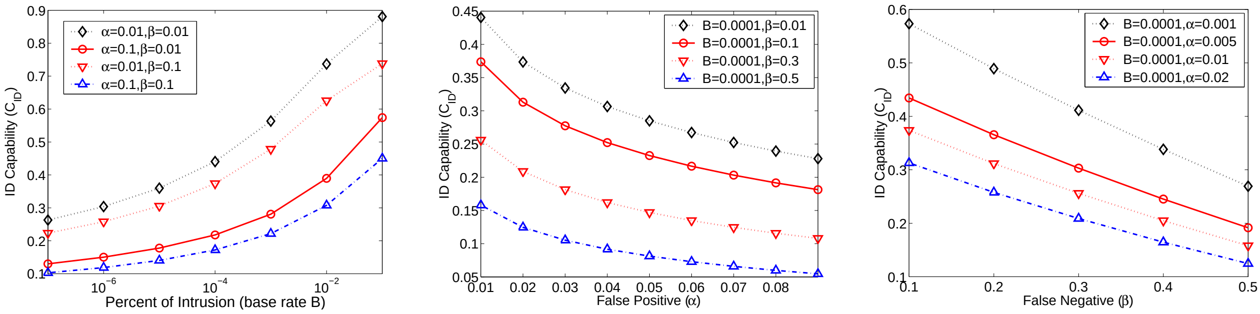
As discussed in Section 2.1, mutual information measures the reduction of uncertainty of the input by knowing the IDS output. We normalize it using the entropy (i.e., the original uncertainty) of the input. Thus, C_{ID} is the ratio of reduction of uncertainty of the IDS input given the IDS output. Its value range is $[0, 1]$. Obviously, a larger C_{ID} value means that the IDS has a better capability in classifying input events correctly.

C_{ID} can also be interpreted in the following way. Consider $\vec{X}$ as a stochastic binary vector which is the "correct assessment" of the input data stream $\vec{S}$, i.e. the correct indication whether each stream unit is an intrusion or not. The detection algorithm is a deterministic function acting on $\vec{S}$, yielding a bitstring $\vec{Y}$ that should ideally be identical to $\vec{X}$. The IDS has to make correct guesses about the unknown $\vec{X}$, based on the input stream $\vec{S}$. The actual number of required binary guesses is $H(\vec{X})$, the "real" information content of $\vec{X}$. Of these, the number correctly guessed by the IDS is $I(\vec{X};\vec{Y})$ (see Figure 2 for the intersection $H(X) \bigwedge H(Y)$). Thus $I(\vec{X};\vec{Y})/H(\vec{X})$ is the fraction of correct guesses.

Using the definitions in Section 2.1 and the abstract model of IDS input (X) and output (Y) as shown in Figure 1, we have $C_{ID} = I(X;Y)/H(X) = (H(X) - H(X|Y))/H(X)$, $H(X) = -\sum_x p(x)\log p(x) = -B\log B - (1-B)\log(1-B)$, and

$$\begin{aligned} H(X|Y) &= -\textstyle\sum_x \sum_y p(x)p(y|x)\log \frac{p(x)p(y|x)}{p(y)} \\ &= -B(1-\beta)\log \frac{B(1-\beta)}{B(1-\beta)+(1-B)\alpha} - B\beta\log\frac{B\beta}{B\beta+(1-B)(1-\alpha)} \\ &\quad -(1-B)(1-\alpha)\log\frac{(1-B)(1-\alpha)}{(1-B)(1-\alpha)+B\beta} - (1-B)\alpha\log\frac{(1-B)\alpha}{(1-B)\alpha+B(1-\beta)} \end{aligned} \tag{3}$$

We can see that C_{ID} is a function of three basic variables: base rate (B), FP (α), and FN (β). When $B = 0$ or $B = 1$ (i.e., the input is 100% normal or 100% intrusion), $H(X) = 0$. We define $C_{ID} = 1$ for these two cases.



(a) Realistic Environment (Base rate B plots in log-scale). Compare the two sets of lines for changes in α against fixed β values. *If the baserate is low*, changes in α are more significant than changes in β.

(b) The Effect of FP (α). For a small base rate B, (fixed in this plot), even tiny changes in α produce large changes in C_{ID}.

(c) The Effect of FN (β). For a small base rate B, (fixed in this plot), only large changes in β have a significant effect on C_{ID}.

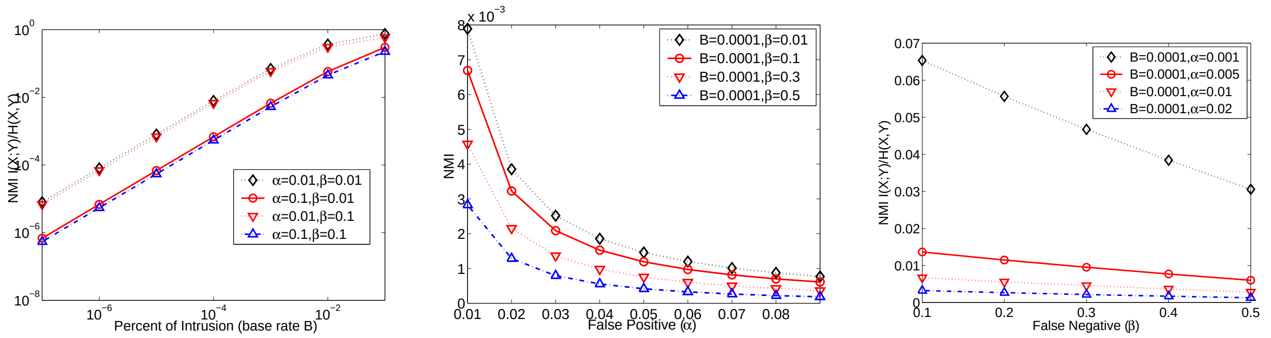
Figure 3: Intrusion Detection Capability. For a realistic low base rate, C_{ID} is more sensitive to changes in α than changes in β. See Appendix A.1 for a formal proof.

From Figure 3(a) we can see the effect of different base rates on C_{ID}. In realistic situations where the base rate (B) is very low, an increase in B will improve C_{ID}. We should emphasize that the base rate is not normally controlled by an IDS. However, it is an important factor when studying intrusion detection capability.

Figure 3(a) clearly shows that for low base rates, it is better to decrease FP than FN in order to achieve a better C_{ID}. For example, suppose we have an IDS with a base rate $B = 10^{-5}$, and a $FP = 0.1$, and $FN = 0.1$. If we decrease the FP from 0.1 to 0.01 (a ten-fold decrease), the C_{ID} moves from 0.1405 to 0.3053. If we instead decrease the FN from 0.1 to 0.01, the C_{ID} only moves from about 0.1405 to 0.1778. Thus, for very low base rates, a reduction in FP yields more improvement in intrusion detection capability than the same reduction in FN. This is intuitive as well, if one realizes that both FN and FP are misclassification errors. When the base rate is low, there are more normal packets that have a chance of being misclassified as FP. Even a large change in FN may not be very beneficial if there are few attack packets at risk for misclassification as FN. A formal proof that C_{ID} is more sensitive to FP than to FN is given in Appendix A.1.

We know that in the perfect case where $FP = FN = 0$, the ID capability is always the same ($C_{ID} = 1$) because the IDS classifies the events without mistake. For realistic (low) base rates, the effects of FP and FN are shown in Figure 3(b) and 3(c). C_{ID} will improve with a decrease of both FP and FN. Note that any reasonable (or "allowable") IDS should have detection rate greater than the false positive rate ($1 - FN > FP$). That is, an IDS should be doing better than random guessing, which has FP=FN=50%. Thus, when $1 - FN < FP$, we define $C_{ID} = 0$.

There do exist several other similar metrics based on normalized mutual information in other research areas. For example, in medical image processing *NMI* (Normalized Mutual Information [PMV03], which is defined as $NMI = (H(X) + H(Y))/H(X,Y)$), is used to compare the similarity of two medical images. In fact $NMI = (H(X)+H(Y))/H(X,Y) = (H(X,Y)+I(X;Y))/H(X,Y) = 1+I(X;Y)/H(X,Y)$. It ranges from 1 to 2. For comparison with C_{ID}, we can plot NMI using $NMI = I(X;Y)/H(X,Y)$ (omitting the "1 plus" from the term as a constant) in Figure 4.



(a) Realistic Environment (plot in log-scale). For the same α values, changes in β produce almost no difference in NMI. Note that the sets of lines almost overlap.

(b) The Effect of FP (α). For a realistic base rate B, the plot of α to NMI is nearly two orders of magnitude less sensitive than C_{ID}. Compare this plot to the C_{ID} plot in Figure 3(b).

(c) The Effect of FN (β). Again, for a realistic base rate B, the plot of β against NMI is nearly one order of magnitude less sensitive than C_{ID}. Compare this plot to the C_{ID} performance in Figure 3(c).

Figure 4: NMI=I(X;Y)/H(X,Y). Using a realistic base rate B, we plot NMI against changes in α and β. Compared to Figure 3, NMI is far less sensitive than C_{ID}. Note the orders of magnitude difference in scales used in this plot, and Figure 3.

We can see from Figure 4(a) that *NMI* has a similar trend as C_{ID}. But we clearly see that NMI is not sensitive to FN in that a variation of FN has almost no effect. For example, when $FP = 0.01$, if we vary FN from 0.01 to 0.1, *NMI* remains almost the same. The reason of this is because in a realistic IDS operation environment, the base rate is very low (close to zero) which means the uncertainty of X is close to zero. Thus, the entropy of X (nearly zero) is far less than the entropy of Y because the IDS can produce many false positives, as shown in the right part of Figure 2. We have $NMI = I(X;Y)/H(X,Y) = I(X;Y)/(H(X) + H(Y) - I(X;Y))$, and $H(Y) \gg H(X) > I(X;Y)$. We also know that a change of FN will only cause a very slight change of $I(X;Y)$. (Recall the discussion above, where a low base rate implies there are few attack packets exposed to the risk of being misclassified as FN). Thus, a change in FN actually has very little effect on the change of NMI.

Further, consider the plots in Figure 3(c) with Figure 4(c). For equivalent ranges of FN, the y-axis for the NMI plot in Figure 4 ranges from 0 to 0.07, while the axis for the C_{ID} ranges from 0.1 to 0.6. Thus, C_{ID} is almost an order of magnitude more sensitive than NMI to changes in FN. Similarly, the corresponding FP plots in Figure 3(b) and Figure 4(b) show that C_{ID} is approximately 100 times more sensitive than NMI to equivalent shifts in FP rates. For all these reasons, NMI is not a good measure of intrusion detection capability.

In other domains, where the relation $H(X) \ll H(Y)$ does not apply, NMI may be a suitable metric.

NMI is a symmetric measure. There is an asymmetric measure called NAMI (Normalized Asymmetric Mutual Information) in [Str02], which is defined as $NAMI = I(X;Y)/H(Y)$. This metric has the same problem as NMI in that it is relatively insensitive to changes in FN. In realistic IDS scenarios, the base rate is low, and $H(X) \ll H(Y)$. Accordingly, $H(Y) \approx H(X,Y)$. So $NAMI \approx NMI$, and is unsuitable for an intrusion detection metric.

3 Analysis and Comparison

In this section we provide an in-depth analysis of existing IDS metrics and compare them with our new metric C_{ID}.

3.1 ROC Curve Based Measurement

An ROC curve shows the relationship between TP and FP but by itself cannot be used to determine the best IDS operation point. ROC curves can sometimes be used for comparing IDSs. If ROC curves of two IDSs do not "cross" (i.e., one is *always* above the other), then the IDS with the top ROC curve is better. However, if the curves do cross, the area under ROC curve (AUC) can be used for comparison. However, this may not be a "fair" comparison because AUC measures all possible operation points of an IDS, while in practice an IDS is fine- tuned to a particular (optimal) configuration (e.g., using a particular threshold).

Gaffney and Ulvila [GU01, UG03] proposed to combine cost-based analysis with ROC to compute an expected cost for each IDS operation point. The expected cost can then be used to select the best operation point and to compare different IDSs. They assigned a cost C_α[1] for responding to a false alarm and cost C_β for every missed attack. They defined the cost ratio as $C = C_\beta/C_\alpha$. Using a decision tree model, the expected cost of operating at a given point on the ROC curve is the sum of the products of the probabilities of the IDS alerts and the expected costs conditional on the alerts. This expected cost is given by the following equation:

$$C_{exp} = Min\{C\beta B, (1-\alpha)(1-B)\} + Min\{C(1-\beta)B, \alpha(1-B)\} \tag{4}$$

In a realistic IDS operation environment, the base rate is very low, say 10^{-5}. The α is also very low, say 10^{-3} (because most IDSs are tuned to have very low α), while β may not be as low, say 10^{-1}. So we can reasonably assume $B < \alpha \ll \beta < 1$. If we have selected a very small C (say, less than $\alpha/(B(1-\beta))$), then

$$C_{exp} = C\beta B + C(1-\beta)B = CB$$

This means that whatever false positive and false negative rates are, the expected cost metric remains the same CB! If we have chosen a very large C (say, larger than $1/B$), then the expected cost will become

$$C_{exp} = (1-\alpha)(1-B) + \alpha(1-B) = 1 - B$$

Again in this case it has nothing to do with α and β.

Consider that $1 - \alpha \approxeq 1 - B \approxeq 1$ in realistic situations, we can approximate Eq.(4) as

$$C_{exp} = Min\{C\beta B, 1\} + Min\{C(1-\beta)B, \alpha\} \tag{5}$$

Above equation can be rewritten as,

$$\begin{array}{rcll} C_{exp} & = & CB & if \quad CB < \frac{\alpha}{1-\beta} \\ & = & C\beta B + \alpha & if \quad \frac{\alpha}{1-\beta} < CB < 1 \\ & = & 1 + \alpha & if \quad CB > 1 \end{array} \tag{6}$$

[1] Note that the notation C_{ID} has no relation to the notation C_α. Our metric measures capability, while Gaffney and Ulvila measured cost.

From the above analysis, we can see that C is a very important factor in determining the expected cost. However, C is not an objective measure. In fact, in practice, it is very hard to determine the appropriate value of C. Furthermore, in [GU01, UG03], Gaffney and Ulvila assumed a stationary cost ratio (C). This may not be appropriate because in practical situations, the relative cost (or tradeoff) of false alarm and missed attack changes as the total number of false alarms and missed attacks changes.

To conclude, using ROC alone has limitations. Combining it with cost analysis can be useful but it involves a subjective parameter that is very hard to estimate because a good (risk) analysis model is hard to obtain in many cases. On the other hand, our C_{ID} is a very natural and objective metric.

3.2 Bayesian Detection Rate

Bayesian detection rate [Axe99] is in fact the positive predictive value (PPV), which is the probability that there is an intrusion when the IDS outputs an alarm. Similarly, Bayesian negative rate (or negative predictive value, NPV) is the probability that there is no intrusion when the IDS does not output an alarm. These metrics are very important from a usability point of view because ultimately, the IDS alarms are useful only if the IDS has high PPV and NPV. Both PPV and NPV depend on TP and FP, and are sensitive to base rate. Thus, they can be expressed using Bayes theorem so that the base rate can be entered as a piece of prior information about the IDS operational environment in the Bayesian equations.

The Bayesian detection rate (PPV) is defined as [Axe99]:

$$P(I|A) = \frac{P(I,A)}{P(A)} = \frac{P(I)P(A|I)}{P(I)P(A|I) + P(\neg I)P(A|\neg I)} = \frac{B(1-\beta)}{B(1-\beta) + (1-B)\alpha}$$

Here, A means IDS outputs an alert and I indicates there is an intrusion. That is, $P(I)$ is the base rate B, $P(A|\neg I)$ is FP or α, $P(\neg A|I)$ is FN or β, and $P(A|I)$ is TP or $1-\beta$.

Similarly, the Bayesian negative rate (NPV) is:

$$P(\neg I|\neg A) = p(X=0|Y=0) = \frac{(1-B)(1-\alpha)}{(1-B)(1-\alpha) + B\beta}$$

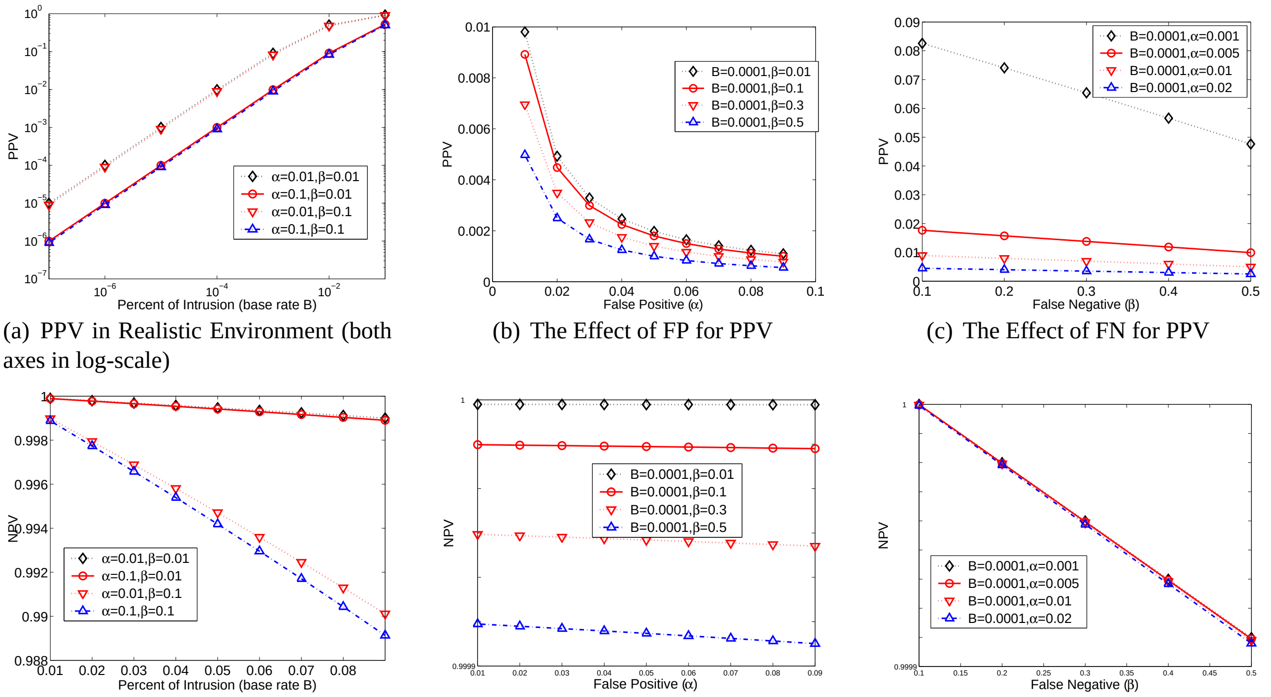
Clearly PPV and NPV are functions on variables B, α, β. Their relationship is shown in Figure 5. We can see that both PPV and NPV will increase if FP and FN decrease. This is intuitive because lower FP and FN should yield better detection results.

Figures 5(a) and 5(b) shows that FP actually dominates PPV when the base rate is very low. This means that in most operation environments (when B is very low), PPV almost totally depends only on FP. It also only changes very slightly with different FN values. For example, when $FP = 0.01$, if we vary FN from 0.01 to 0.1, PPV remains almost the same. This shows that PPV is not sensitive to FN. Figure 5(c) shows PPV is not as sensitive to FN as C_{ID}. Similarly, Figures 5(d), 5(e), and 5(f) show that NPV is not sensitive to FP and FN.

To conclude, when evaluating IDS from an usability point of view, both PPV and NPV are needed. However, similar to the situation with TP and FP, there is no existing objective method to integrate these metrics. On the other hand, we can rewrite Equation (3) as

$$H(X|Y) = -B(1-\beta)\log PPV - B\beta\log(1-NPV) - (1-B)(1-\alpha)\log NPV - (1-B)\alpha\log(1-PPV)$$

We can see that our new metric C_{ID} has incorporated both PPV and NPV in measuring intrusion detection capability. C_{ID} in fact unifies all existing commonly used metrics, i.e., TP, FP, PPV, and NPV. It also factors in base rate, a measure of the IDS operation environment.



(a) PPV in Realistic Environment (both axes in log-scale)

(b) The Effect of FP for PPV

(c) The Effect of FN for PPV

(d) NPV in Realistic Environment

(e) The Effect of FP for NPV

(f) The Effect of FN for NPV

Figure 5: Positive and Negative Predictive Value. These plots, similar to those in Figures 4 show that PPV and NPV are not sensitive measures when the base rate is low. In (a), changes in β (for the same α values) have nearly no effect on PPV. In (b) for a low base rate, changes in α have a small effect on PPV. The insensitivity of PPV is also seen in (c), where changes in β do not result in large changes in PPV. The same is true for NPV, in graphs (d), (e), and (f), which show that changes in α and β do not significantly affect NPV.

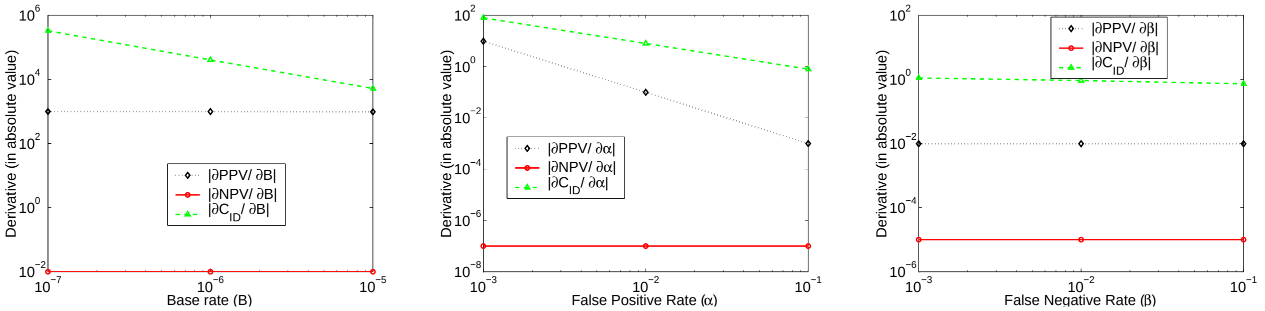
3.3 Sensitivity Analysis

We already see one important advantage of C_{ID} over existing metrics: it is a single unified metric, and is very intuitive and appealing, with a grounding in information theory.

In this section we analyze in depth why C_{ID} is more sensitive than traditional measures in realistic situations (i.e., where the base rate is low). IDS design and deployment often results in slight changes in these parameters. For example, when fine-tuning an IDS (e.g., setting a threshold), different operation points have different TP and FP. Being sensitive means that C_{ID} can be used to measure even the slight improvements to an IDS. PPV and NPV, on the other hand, require more dramatic improvements to an IDS to yield measurable differences. Similarly, C_{ID} provides a fairer comparison of two IDSs because, for example, a slightly better FN actually shows more of an improvement in capability than in PPV. In short, C_{ID} is a more "precise" metric.

As we know, the scales of PPV, NPV, C_{ID} are all the same, i.e., from 0 to 1. This provides a fair situation to test their sensitivity. To investigate how much more sensitive C_{ID} is compared to PPV and NPV, we can perform a differential analysis of base rate B, false positive FP, and false negatives FN to study the effect of changing these parameters on PPV, NPV, and C_{ID}. We can assume that $B \ll 1$ and $\alpha \ll 1$, i.e., for most IDSs and their operation environments, base rate and false positive rates are very low. Approximate derivatives are shown below. (The detailed steps in this differential analysis appear in Appendix A.2). Note that although we originally plot Figure 6 according to Equation(7) where we simplify $B \ll 1$ and $\alpha \ll 1$, it turns out we will get almost the same figures when we do the numerical solution on the differential formula of PPV, NPV, C_{ID} without any simplification on B, α.

$$\begin{aligned}
&\frac{\partial}{\partial B}PPV \approx +\frac{\alpha(1-\beta)}{(\alpha+(1-\beta)B)^2},\ \frac{\partial}{\partial \alpha}PPV \approx -\frac{(1-\beta)B}{(\alpha+(1-\beta)B)^2},\ \frac{\partial}{\partial \beta}PPV \approx -\frac{\alpha B}{(\alpha+(1-\beta)B)^2}\\
&\frac{\partial}{\partial B}NPV \approx -\beta,\ \frac{\partial}{\partial \alpha}NPV \approx -B\beta,\ \frac{\partial}{\partial \beta}NPV \approx -B\\
&\frac{\partial}{\partial B}C_{ID} \approx \frac{1}{H(X)^2}(\alpha(\log\frac{\alpha}{\alpha+(1-\beta)B} - B\beta)\log B + B(\log\frac{(1-\beta)B}{\alpha+(1-\beta)B} - \beta\log\frac{1-\beta}{\beta(\alpha+(1-\beta)B)}))\\
&\frac{\partial}{\partial \alpha}C_{ID} \approx \frac{1}{H(X)}\log\frac{\alpha}{\alpha+(1-\beta)B},\ \frac{\partial}{\partial \beta}C_{ID} \approx -\frac{B}{H(X)}\log\frac{1-\beta}{\beta(\alpha+(1-\beta)B)}
\end{aligned} \tag{7}$$



(a) Dependence on base rate analysis ($\alpha = 0.001, \beta = 0.01$)

(b) Dependence on false positive rate analysis ($B = 0.00001, \beta = 0.01$)

(c) Dependence on false negative rate analysis ($B = 0.00001, \alpha = 0.001$)

Figure 6: Derivative analysis (in absolute value). In every situation C_{ID} has the highest sensitivity, compared to PPV and NPV. For realistic situations, its derivative is always higher than other measures.

Figure 6 shows the derivatives (in absolute value) for different metrics. We only need to see the absolute value of derivative. A larger derivative value shows more sensitivity to changes. For example, in Figure 6(a), a change in the base rate results in very tiny change in NPV. PPV improves with the change, but not as much as C_{ID}. Clearly, from Figure 6 we can see that C_{ID} is more sensitive to changes in B, FP, FN than PPV and NPV.

For small base rates and false negative rates, PPV is more sensitive to changes in the base rate [Axe99], than changes in FP. It is least sensitive to FN. Given the same base rate and FP, the change of FN has a very small effect on PPV. This implies that for a large difference in FN but a small difference in FP, the IDS with the smaller FP will have a better PPV. For example, suppose we have two IDSs with the same base rate $B = 0.00001$, IDS_1 has $FP = 0.2\%$, $FN = 1\%$ while IDS_2 has $FP = 0.1\%$, $FN = 30\%$. Although IDS_1 has a far lower FN ($1\% \ll 30\%$) and slightly higher FP ($0.2\% > 0.1\%$), its PPV (0.0049) is still lower than IDS_2 (0.007). On the other hand, its C_{ID} (0.4870) is greater than IDS_2 (0.3374).

NPV on the other hand is more sensitive to B and FN. It does not change much for a change in FP. This implies that for large difference in FP but small difference in FN, the one with the smaller FN will have a better NPV. For example, two IDS's with the same base rate 0.00001, IDS_1 has $FP = 0.1\%$, $FN = 2\%$ while IDS_2 has $FP = 2\%$, $FN = 1\%$. Although IDS_1 has far lower FP ($0.1\% \ll 2\%$) and slightly higher FN ($2\% > 1\%$), its NPV (1-2.002e-6) is still lower than IDS_2 (1-1.0204e-6). On the other hand, its ID Capability (0.4014) is greater than IDS_2 (0.2555).

To conclude, C_{ID} is a more precise and sensitive measure than PPV and NPV.

4 Performance Measurement Using C_{ID}

4.1 Selection of Optimal Operating Point

C_{ID} factors in all existing measurements, i.e., base rate, FP, FN, PPV, and NPV, and is the appropriate performance measure to maximize when fine tuning an IDS (so as to select the best IDS operation point). The thus obtained operation point is the best that can be achieved by the IDS in terms of its intrinsic ability to classify input data. For anomaly detection systems, we can change some threshold in the detection algorithm so that we can achieve different corresponding FP and FN, and create an ROC curve. In order to obtain the best optimized operational point, we can calculate a corresponding ID capability for every point in the ROC. We then select the point which gives the highest ID capability, and the threshold corresponding to this point provides the optimal threshold for use in practice.

To illustrate, we first give a numerical example. We take the two ROC examples from [GU01]. These two intrusion detectors, denoted as IDS_1 and IDS_2, have ROC curves that were determined from data in the 1998 DARPA off-line intrusion detection evaluation [GLC$^+$]. We do not address how these ROC curves were obtained, and instead merely use them to demonstrate how one selects an optimized operating point using C_{ID}.

As in [GU01], the IDS_1 ROC can be approximated as $1-\beta = 0.6909 \times (1 - exp(-65625.64\alpha^{1.19}))$. The IDS_2 ROC is approximated as $1-\beta = 0.4909 \times (1 - exp(-11932.6\alpha^{1.19}))$. For both IDSs, the base rate is $B = 43/660000 = 6.52 \times 10^{-5}$. From these two ROC curves, we can get their corresponding ID Capability curves in Figure 7.

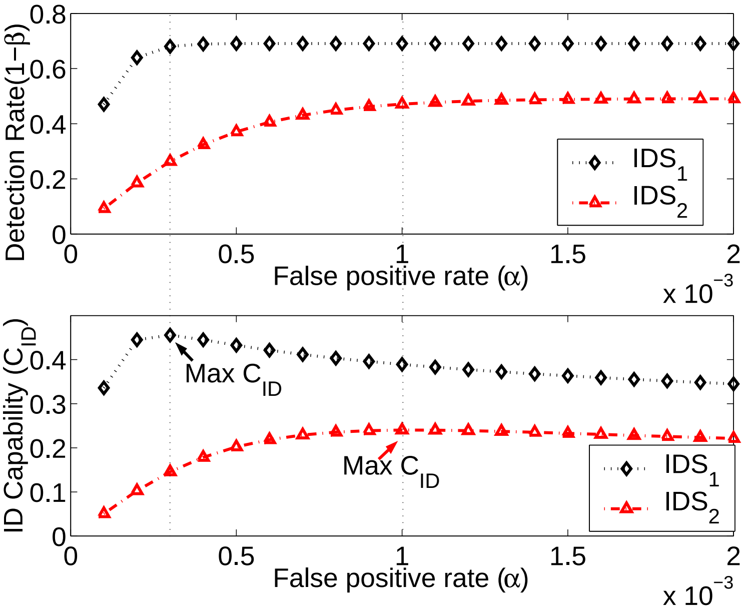


Figure 7: IDS_1 and IDS_2 ROC curves and corresponding ID Capability Curves. These plots, based on values reported in [GU01], show how C_{ID} can be used to select an optimal operating point. It is not clear how simple ROC analysis could arrive at this same threshold.

We can see that IDS_1 achieves the best ID Capability (0.4557) when the false positive rate is approximately 0.0003 (corresponding to detection rate $1-\beta = 0.6807$). So this point (with the corresponding threshold) provides the best optimized operating point for IDS_1. The optimized operating point for IDS_2 is approximately $\alpha = 0.001, 1-\beta = 0.4711$ and the corresponding maximized ID capability is 0.2403. Thus, to set the optimized threshold one merely has to calculate a C_{ID} for each known point (for its TP and FP) on the ROC curve, and then select the maximum.

4.2 Comparison of Different IDSs

When we get the maximized C_{ID} for every IDS, we can compare their C_{ID} to tell which IDS has better Intrusion Detection Capability. For example, in the previous section, clearly IDS_1 is better than IDS_2 because it has a higher C_{ID}. Granted, in this case it is easy to compare IDS_1 and IDS_2 just from ROC curves. But in many cases comparing ROC curves is not straight forward, particularly when the curves cross.

Consider another simple numerical example with the data taken from [LFe00]. We compare two IDSs which have only *single* point ROC curves (for PROBE attacks). IDS_1 has $FP = 1/660,000$, $TP = 0.88$, while IDS_2 has $FP = 7/660,000$, $TP = 0.97$. The base rate here is $B = 17/(17 + 660,000)$. We note these single point curves were critiqued in [McH00], but here we use it merely as a simple numerical example of how C_{ID} might compare two IDSs. IDS_1 has $C_{ID} = 0.8390$ and IDS_2 has $C_{ID} = 0.8881$. Thus, IDS_2 is a little better than IDS_1. Reaching this same conclusion using just the ROC curves in [LFe00] is not obvious.

The relative C_{ID} between different IDSs is fairly stable even if the base rate in realistic situations may change a little. This can be easily seen from Figure 3(a). The four curves do not intersect within the range of base rate from 10^{-7} to 10^{-1}.

4.3 Experiments

To demonstrate how to use the sensitivity of the ID capability measurement to select the optimal operation point (or fine tune an IDS) in practice, we examined several existing anomaly detection systems, and measured their accuracy, C_{ID}, under various configurations. Specifically, we used two anomaly network intrusion detection systems, Packet Header Anomaly Detection (PHAD) [MC01] and Payload Anomaly Detection (PAYL) [WS04]. To demonstrate how to compare two different IDSs using C_{ID}, we compared an anomaly detection system PAYL with another open source signature-based IDS, Snort [Roe99], in terms of their capabilities to detect Web attacks based on the same testing data set.

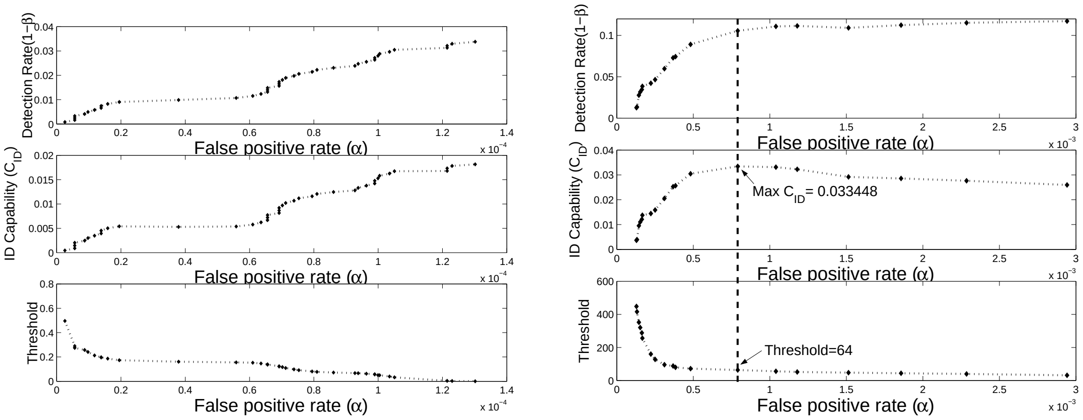
PHAD and PAYL both detect anomalies at the packet level, with PHAD focusing on the packet header, and PAYL using byte frequencies in the payload. We tested PHAD using the DARPA 1999 test data set [LL01], using week 3 for training and weeks 4 and 5 for testing. We configured PHAD to monitor only HTTP traffic. As noted in [WS04], it is difficult to find sufficient data in the DARPA 1999 data set to thoroughly test PAYL, so we used GTTrace, a backbone capture from Georgia Tech network. The GTTrace data set consists of approximately 6 hours of HTTP traffic captured on a very busy 1Gb/s backbone, or approximately 1G of data. We filtered the GTTrace set to remove known attacks, split the trace into training and testing sets, and injected numerous HTTP attacks into the testing set, using tools such as libwhisker [Pup04]. We used C_{ID} to identify an optimal setting for each IDS.

In PHAD, a score is computed based on selected fields in each packet header. If this score exceeds a threshold, then an intrusion alert is issued. Adjusting this threshold yields different TP, FP values, shown in Figure 8(a). We configured PHAD to recognize the attack packets, instead of the attack instances reported in [MC01].

We can see in Figure 8(a), that the C_{ID} curve almost follows the ROC curve (both like straight lines). The reason is that with the DARPA data set, we found the false positive rate for PHAD was fairly low, while the false negative rate was extremely high, with $\beta \approx 1$. As shown in Appendix A.3, given small values of α and large values of β, ROC and C_{ID} can both be approximated as straight lines, and the equation for C_{ID} becomes essentially $K(1 - \beta)/H(X)$, where K is a constant. We note that the authors in [MC01] used PHAD to monitor traffic of all types, and the details of training and testing were also different from our experiments. In particular, we configured PHAD to report each packet involved in an attack instead of reporting the attack instance. Therefore, our PHAD has a high β than reported in [MC01].

One can argue that just selecting the point from the ROC with the highest detection rate is an adequate way to tune an IDS. This may be true in anecdotal cases, as illustrated by our configuration of PHAD. But it is not always the case, as shown in other situations such as Figure 7.

Our reanalysis of PHAD therefore illustrates a worst-case scenario for C_{ID}. With $\beta \approx 1$, and $\alpha \approx 0$, C_{ID} identifies an operating point no better than existing detection measurements, e.g. ROC. Note, however, that C_{ID}



(a) PHAD

(b) PAYL

Figure 8: Experiment Results. a) A low false positive and high false positive rate in the PHAD test means C_{ID} is no better (and no worse) than ROC. b) C_{ID} identifies an optimal threshold in PAYL. A simple ROC analysis would fail to find this point because of an increasing detection rate.

will never return a worse operating point.

In other situations, C_{ID} will outperform existing metrics. Indeed, our analysis of PAYL and the GTTrace data set illustrates a situation where C_{ID} provides a better measure than simple ROC analysis. PAYL requires the user to select an optimal threshold for determining whether observed byte frequencies vary significantly from a trained model. For example, a threshold of 256 allows each character in an observed payload to vary within one standard deviation of the model [WS04]. As before, we can experiment with different threshold values, and measure the resulting FP, FN rates. In Figure 8(b), we see that for the GTTrace data, as the threshold drops, C_{ID} reaches a peak and then drops, while the ROC curves (shown in the top graph) continues to slowly increase.

An analyst using just the top graph in Figure 8(b) might be tempted to set a threshold lower than, say, 8 (where $\alpha = 3 \times 10^{-3}$), because the detection capability still increases, even if the false positive rate grows slightly as well. But using C_{ID}, we see the detection capability actually declines after $C_{ID} = 0.033448$ (marked in Figure 8(b) with a vertical line). Thus, C_{ID} identifies a higher, but optimal operating threshold of 64 (where $\alpha = 0.7 \times 10^{-3}, 1 - \beta = 0.10563$). In this situation, C_{ID} provides a better operating point. It is not obvious how ROC analysis could provide the same optimal threshold.

To demonstrate how C_{ID} can be used to compare different IDS, we ran Snort (Version 2.1.0 Build 9) on the same data as PAYL to compare their capabilities. Since the use of libwhisker which tried to evade snort, we have a poor detection rate with $1 - \beta = 0.0117$ (worse than PAYL), a good false positive rate $\alpha = 0.0000006701$ (better than PAYL). Without C_{ID}, we cannot tell which IDS is better based on existing metrics. With the base rate $B = 0.000010191$, we can calculate $C_{ID} = 0.0081$ for Snort in this testing data. Clearly $0.033448 > 0.0081$, so (optimally configured) PAYL performs better than Snort based on our test data.

Again we emphasize that as with all evaluation attempts, the above results are very related to the testing data in use.

5 Discussion

5.1 Estimation of Base Rate, FP, FN

When we are evaluating IDSs, we should have the evaluation data set where we know the details about the ground truth, i.e., what data are attacks and what data are normal traffic. Thus, we can easily find out the base rate which is the fraction of the attacks in the whole data set. After the testing of IDS on this evaluation data set, we compare the IDS alerts with the ground truth, then we can calculate the false positive rate (the fraction of misclassified normal data in the whole normal data) and false negative rate (the fraction of undetected attacks among all the attack data). Using these basic metrics as inputs, we can finally compute C_{ID}.

If the testing data is a very representative sample of the real situation, we can use the FP, FN in the testing data to approximate the real situation. In machine learning, there is a similar problem: based on an observed error (FP and FN) over a sample of data, how well can we estimate the true error. Using statistical theory, machine learning researchers have already answered this question [Mit97]. Given an observed sample error e_s, with approximately $N\%$ (e.g. 99%) probability, the true error e_t will lies in the interval $e_s \pm z_N\sqrt{\frac{e_s(1-e_s)}{n}}$, where n is the number of records in sample data, z_N is a constant related to the confidence interval $N\%$ we want to reach. For example, if we want approximate 99% confidence intervals then $z_N = 2.58$.

Once we have estimated the FP (α), FN (β), we can also approximately estimate the base rate in real traffic using the estimated FP, FN. First we can easily calculate the alert rate (A_r) of the IDS by dividing the total number of alerts by the total number of data packets. On the other hand, the current alert rate can also be calculated using base rate, α, and β as shown in equation (8)

$$A_r = B(1-\beta) + (1-B)\alpha \tag{8}$$

. Solving above equation for B gives us,

$$B = \frac{A_r - \alpha}{1-\beta-\alpha} \tag{9}$$

. This provides a good estimation of the real world base rate.

5.2 Unit of Analysis

An important problem in IDS evaluation is “unit of analysis” [McH00]. As we have mentioned when introducing the abstract IDS model, for network based intrusion detection, there are at least two units of analysis in different IDSs. Some IDSs (such as Snort, PAYL [WS04], etc) analyze packets and output alerts on the packets. While other IDSs such as Bro analyze traffic based on flows.

Different unit of analysis will result in different base rate even on the same evaluation data set. It does not affect the usage of C_{ID} in fine-tuning an IDS to get optimal operation point. But when comparing different IDSs, we do need to consider this problem. In this paper, we are not trying to solve the “unit of analysis” problem, because this is not a problem specific to C_{ID}, but is a general problem for all the existing evaluation metrics, e.g. TP, FP, etc. We recommend that, in order to provide a fair and meaningful comparison, it is better to run the IDSs based on the same unit of analysis, as well as the same data set and the same detection spaces (or attack coverages).

The “unit of analysis” problem is a general and hard problem in IDS evaluation, regardless of the metrics being used. In some cases, we can also convert the different units to the same one when comparing different IDSs. For example, we can convert a packet-level analysis to a flow-level analysis by defining a flow is malicious when it contains any malicious packet and otherwise it is a normal flow. Using such a conversion, it makes the comparison between a packet-level IDS and a flow-level IDS possible based on the same (“virtual”) granularity or unit of analysis. But this kind of conversion does not always work, especially when the two units are totally un-related, such as packet sequence and system call sequence.

5.3 Involving Cost Analysis in C_{ID}

We have showed that C_{ID} is a very natural and objective metric to measure the intrusion detection capability. But in some cases we may want to include some subjective cost analysis especially when a good risk analysis model is available. For instance, considering a military site that is highly cautious, false negatives are probably the biggest worry. Accordingly, an IDS is considered to be optimal if it minimizes false negatives, even when that implies many false positives (e.g., the military site has got the resources to sort through the false positives). Conversely, a site with a single overloaded operator (or with an automated response system) will likely prefer a low false positive rate because the operator will not use the IDS otherwise (and the response system will cause a lot of damage, respectively). These two examples illustrate the cases where the risk model is clear and cost analysis combing ROC is possibly useful when evaluating IDSs.

Although so far we have only shown the objective property of C_{ID} to measure the natural capability of intrusion detection, we can easily involve cost analysis in C_{ID} as an extension. A possible solution is achieved by using a weighted conditional entropy $H(X|Y)$ when calculating $C_{ID} = (H(X) - H(X|Y))/H(X)$. We can slightly change the original form of conditional entropy and place weights in. Now

$$H_w(X|Y) = \frac{-\sum_x \sum_y w_{xy} p(x,y) \log p(x|y)}{\sum_x \sum_y w_{xy}}$$

where w_{xy} means the weight/cost considered when $X = x, Y = y$. We can set larger weight of w_{xy} when we believe the situation $X = x, Y = y$ costs more. For instance, in the military network example, we can define a very large weight on w_{10} which essentially gives more weight to missed attacks ($X = 1$ while $Y = 0$), i.e., false negative. In this weighted setting, C_{ID} will give more preference to FN than FP. Similarly, we can set a larger weight of w_{01} in the case with a single overloaded operator (or with an automated response system), which means false positive ($X = 0, Y = 1$) is more important in the analysis. In such a cost-based extension, C_{ID} can achieve similar capability as ROC combining cost analysis. Note that there are some other possible cost-based extensions on C_{ID}. We will further study this topic in our future work.

6 Related Work

Intrusion detection has been a field of active research for more than two decades, and many IDSs have been developed (a good survey is by Debar et al. [DDW99]). Since our work is concerned with a fundamental problem in intrusion detection, we first discuss some relevant fundamental (theoretical) research in this field.

Denning [Den87] introduced an intrusion detection model and proposed a framework for a general-purpose intrusion-detection expert system. Several statistical models were proposed to build normal profiles. This is the first systematic work in IDS.

Helman and Liepins [HL93] modeled the normal traffic and attack traffic as the output of two independent stationary stochastic processes. Assuming the knowledge of the base rate and exact distributions of the normal and attack traffic, their algorithm can produce the misuse detector with the lowest Bayesian error rate. In the absence of base rate, they proposed a metric called prioritization error to determine the accuracy of the intrusion detector. They also proved that for a realistic environment where one does not know the exact distribution of normal and attack traffic, it is very hard to optimize the IDS for an error rate. This motivated a heuristic approach; however, the authors did not propose any solution for the evaluation of IDS in such an environment.

Axelsson [Axe00] argued that the well established signal detection and estimation theory bears similarities with the IDS domain. We can think of the anomaly detection model as "signal source" in detection and estimation theory, the auditing mechanism as "signal transmission", the audit data as "observation space", and in both cases, the task is to derive detection rules. Thus, results from detection and estimation theory, which have been found applicable to a wide range of problems, may be used in the IDS research. However this work is very preliminary and it is unclear how these similarities can benefit the design and evaluation of IDS in practice.

Maxion et al. [MT00] studied the relationship between data regularity and anomaly detection performance. The study focused on sequence data, and hence regularity is defined as conditional entropy. The key result from experiments on synthetic data is that when an anomaly detection model is tested on datasets with varying regularity values, the detection performance also varies. Lee et al. [LX01] applied information theoretic measurement to describe the characteristics of audit data set, suggest the appropriate anomaly detection model, and explain the performance of the models. Our work is another application of information theory to IDS and provides a natural and unified metric of intrusion detection capability.

In the area of IDS evaluation, true positive rate and false positive rate are two commonly used metrics. To consider both of these metrics, we can ROC (receiver operating characteristic) curve [HW66] based analysis, which has been already well studied in other fields such as medical diagnostic tests [Swe88]. Lippmann et al. [LFe00] evaluated IDSs on the 1998 DARPA Intrusion Detection Evaluation Data Set and used ROC curves to evaluate (and implicitly compare) the IDSs. McHugh [McH00] pointed out that the evaluation in [LFe00] had serious shortcomings. For example, the appropriateness of ROC analysis is very questionable when the IDSs only produce 0 or 1 outputs and the proper unit of analysis and measurement is different for different detectors. McHugh also called for the more helpful measure of IDS performance. Our work is an attempt to develop a better metric. Gaffney and Ulvila [GU01, UG03] combined ROC curves with cost analysis methods to compute the expected cost for an IDS so that different IDSs can be evaluated and compared based on their expected costs. This approach is not practical because the result depends on the subjective estimate of the cost ratio between true and false positives.

Axelsson [Axe99] proposed two other metrics, the Bayesian detection rate and the Bayesian negative rate. These are in fact the Bayesian representations of positive predictive value (PPV) and negative predictive value (NPV), which are commonly used in medical diagnostic [Swe88]. Axelsson's main conclusion is that given that the base rate is very low in most environments, the false alarm rate needs to be a lot lower than what most current algorithms can achieve in order to have a reasonable Bayesian detection rate.

The existing metrics are all useful. However, the lack of a unified metric makes it hard to fine tune and evaluate an IDS. Our new metric, Intrusion Detection Capability, is derived from analyzing the intrusion detection process from an information-theoretic point of view. It nicely unifies all the existing objective measures of IDS detection capability.

RIDAX [Dac] method and tool, developed by the IBM Zurich team in the context of the European MAFTIA project, also noticed the fact that a mere counting of false and true positives is insufficient. They proposed a set of metrics like precision, recall, etc., as used in the information-retrieval field. Their approach is very different from C_{ID} because they focus on assessing the completeness and utility of arbitrary IDS combinations, while we try to capture the intrinsic capability of IDS using an information-theoretic approach.

Our new metric is similar to but different from NMI (Normalized Mutual Information, $(H(A) + H(B))/H(A, B)$) used in medical image registration [PMV03] and NAMI (Normalized Asymmetric Mutual Information, $I(X; Y)/H(Y)$) [Str02]. These other metrics are not as sensitive as C_{ID} for realistic intrusion detection scenarios, as discussed in Section 2.

7 Conclusion and Future Work

The contributions of this paper are both theoretical and practical.

We provided an in-depth analysis of existing IDS metrics. We argued that the lack of a unified metric makes it hard to fine-tune an IDS and compare different IDSs. We studied the intrusion detection process from the viewpoint of information theory, and proposed a natural, unified metric to measure the capability of IDS in terms of its capability to correctly classify input events. Our metric, Intrusion Detection Capability, or C_{ID}, is simply the ratio of mutual information between the IDS input and output, and entropy of IDS input. This intuitive metric combines all commonly used metrics, i.e., true positive rate, false positive rate, and both positive and negative predictive values. It also factors in base rate, a important measure of the IDS operation environment.

Using this metric, one can choose the best (optimized) operation point of an IDS (e.g., the threshold for an anomaly detection system). Further, since C_{ID} is normalized, we can compare different ID models, even though their FP, FN rates are different. We presented numerical experiments and case studies.

This paper has not presented every application for Intrusion Detection Capability, and numerous extensions are possible. For example, the abstract model for intrusion detection, presented in Section 2, can be easily expanded to multiple IDS models. The normalized value of C_{ID} also allows designers to compare different IDS configurations, topologies, and deployment strategies, even if individual components have different FP and FN rates.

An obvious extension of C_{ID} comes from rethinking the simple model of IDS inputs and outputs, X, Y, represented as a 1 or 0. We can instead encode different types of attacks into X, Y, creating a more accurate model, especially in the context of signature-based IDS. We are aware of this more accurate model and will use C_{ID} to measure more signature-based detection systems. We believe C_{ID} is a useful yardstick for the type of evaluation and testing of misuse detection systems performed by Vigna et al. [VRB04]. We will publish our result in the near future.

We notice that our abstract model for the intrusion detection process can be further studied using channel capacity models from information theory. Multiple processes (or layers) of IDS can be considered as multiple (chained) channels. We can analyze and improve both internal and external designs of IDS, instead of only considering the intrusion detection process as a whole black box.

References

[Axe99] S. Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *Proceedings of ACM CCS'1999*, November 1999.

[Axe00] Stefan Axelsson. A preliminary attempt to apply detection and estimation theory to intrusion detection. Technical Report 00-4, Dept. of Computer Engineering, Chalmers Univerity of Technology, Sweden, March 2000.

[CT91] Thomas Cover and Joy Thomas. *Elements of Information Theory*. John Wiley, 1991.

[Dac] M. Dacier. Design of an intrusion-tolerant intrusion detection system, Maftia Project, deliverable 10. Available at http://www.maftia.org/deliverables/D10.pdf. 2005.

[DDW99] Herve' Debar, Marc Dacier, and Andreas Wespi. Towards a taxonomy of intrusion-detection systems. *Computer Networks*, 31(8):805–822, 1999.

[Den87] D. Denning. An intrusion-detection model. *IEEE Transactions on Software Engineering*, 13(2), Feb 1987.

[GLC$^+$] I. Graf, R. Lippmann, R. Cunningham, K. Kendall D. Fried, S. Webster, and M. Zissman. Results of DARPA 1998 off-line intrusion detection evaluation. Presented at DARPA PI Meeting, 15 December 1998.

[GU01] John E. Gaffney and Jacob W. Ulvila. Evaluation of intrusion detectors: A decision theory approach. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, May 2001.

[HL93] P. Helman and G. Liepins. Statistical foundations of audit trail analysis for the detection of computer misuse. *IEEE Transactions on Software Engineering*, 19(9), September 1993.

[HW66] J. Hancock and P. Wintz. *Signal Detection Theory*. McGraw-Hill, 1966.

[LFe00] R. P. Lippmann, D. J. Fried, and I. Graf etc. Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation. In *Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX'00)*, 2000.

[LL01] Massachusetts Institute of Technology Lincoln Laboratory. 1999 darpa intrusion detection evaluation data set overview. http://www.ll.mit.edu/IST/ideval/, 2001.

[LX01] W. Lee and D. Xiang. Information-theoretic measures for anomaly detection. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, May 2001.

[MC01] Matthew V. Mahoney and Philip K. Chan. Phad: Packet header anomaly detection for indentifying hostile network traffic. Technical Report technical report CS-2001-4, Florida Tech, 2001.

[McH00] John McHugh. Testing intrusion detection systems: A critique of the 1998 and 1999 darpa off-line intrusion detection system evaluation as performed by lincoln laboratory. *ACM Transactions on Information and System Security*, 3(4), November 2000.

[Mit97] Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[MT00] R. Maxion and K. M. C Tan. Benchmarking anomaly-based detection systems. In *Proceedings of the 1st International Conference on Dependable Systems and Networks (DSN'00)*, 2000.

[Pax99] Vern Paxson. Bro: A system for detecting network intruders in real-time. *Computer Networks*, 31(23-24):2435–2463, December 1999.

[PMV03] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual information based registration of medical images: A survey. *IEEE Trans on Medical Imaging*, 22(8):986–1004, Aug 2003.

[PN98] T. H. Ptacek and T. N. Newsham. Insertion, evasion, and denial of service: Eluding network intrusion detection. Technical report, Secure Networks Inc., January 1998. http://www.aciri.org/vern/Ptacek-Newsham-Evasion-98.ps.

[Pup04] Rain Forest Puppy. Libwhisker official release v2.1, 2004. Available at http://www.wiretrip.net/rfp/lw.asp.

[PZC+96] Nicholas J. Puketza, Kui Zhang, Mandy Chung, Biswanath Mukherjee, and Ronald A. Olsson. A methodology for testing intrusion detection systems. *IEEE Transactions on Software Engineering*, 22(10):719–729, 1996.

[Roe99] M. Roesch. Snort - lightweight intrusion detection for networks. In *Proceedings of USENIX LISA'99*, 1999.

[Str02] Alexander Strehl. Relationship-based clustering and cluster ensembles for high-dimensional data mining, May 2002. PhD thesis, The University of Texas at Austin.

[Swe88] John A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

[UG03] Jacob W. Ulvila and John E. Gaffney. Evaluation of intrusion detection systems. *Journal of Research of the National Institute of Standards and Technology*, 108(6), November-December 2003.

[VRB04] G. Vigna, W. Robertson, and D. Balzarotti. Testing network-based intrusion detection signatures using mutant exploits. In *Proceedings of the ACM Conference on Computer and Communication Security (CCS'04)*, 2004.

[WS04] Ke Wang and Salvatore J. Stolfo. Anomalous payload-based network intrusion detection. In *Proceedings of RAID'2004*, September 2004.

A Appendix

A.1 Proof that C_{ID} is more sensitive to FP than to FN

This is equivalent to prove $\left|\frac{\partial C_{\mathrm{ID}}}{\partial \alpha}\right| > \left|\frac{\partial C_{\mathrm{ID}}}{\partial \beta}\right|$. The derivatives of C_{ID} are given by

$$\frac{\partial C_{\mathrm{ID}}}{\partial \alpha} = \frac{1-B}{h(B)} \log \frac{\alpha}{1-\alpha}\frac{1-q}{q} = \frac{-1}{1-\log B}\left[\frac{1-\beta}{\alpha} - \frac{\beta}{1-\alpha} + \mathcal{O}(B)\right]$$

$$\frac{\partial C_{\mathrm{ID}}}{\partial \beta} = \frac{B}{h(B)} \log \frac{\beta}{1-\beta}\frac{q}{1-q} = \frac{-1}{1-\log B}\left[\log \frac{1-\beta}{\alpha}\frac{1-\alpha}{\beta} + \mathcal{O}(B)\right]$$

where $q = P(Y=1)$, h(.) is defined as $h(x) = -xlog(x) - (1-x)log(1-x)$. The rightmost parts of the two equations are based on a Taylor expansion.

We define a monotonically increasing function $f(z) = z - \log z$ and two constants $z_1 = (1-\beta)/\alpha$ and $z_2 = (1-\alpha)/\beta$. Using these definitions and derivatives, we can write

$$\left|\frac{\partial C_{\mathrm{ID}}}{\partial \alpha}\right| - \left|\frac{\partial C_{\mathrm{ID}}}{\partial \beta}\right| = \frac{f(z_1) - f(1/z_2) + \mathcal{O}(B)}{1 - \log B}. \tag{10}$$

B is very close to zero in realistic IDS situation. We will neglect the $\mathcal{O}(B)$ terms. In the ordinary operational regime one has $1-\beta > \alpha$ (and hence $\beta < 1-\alpha$). Thus $z_1 > z_2 > 1$. We use the fact that f is an increasing function and that $z_1 > z_2$ to write $f(z_1) > f(z_2)$. Then we use the property that $f(z) - f(1/z) > 0$ for $z > 1$. Since $z_2 > 1$, we have $f(z_2) - f(1/z_2) > 0$. We conclude that $|\partial C_{\mathrm{ID}}/\partial\alpha| - |\partial C_{\mathrm{ID}}/\partial\beta| > 0$.

A.2 Partial Differential Analysis

For approximation purpose, we use $1 - B \approx 1$; $1-\alpha \approx 1$; $\log(1-B) \approx -B$; $\log(1-\alpha) \approx -\alpha$.

Equations below show the partial derivatives of PPV.

$$\begin{aligned}
\tfrac{\partial}{\partial B} PPV &= +\tfrac{(1-\beta)\alpha}{(B(1-\beta)+(1-B)\alpha)^2} \approx +\tfrac{\alpha(1-\beta)}{(\alpha+(1-\beta)B)^2} \\
\tfrac{\partial}{\partial \alpha} PPV &= -\tfrac{B(1-\beta)(1-B)}{(B(1-\beta)+(1-B)\alpha)^2} \approx -\tfrac{(1-\beta)B}{(\alpha+(1-\beta)B)^2} \\
\tfrac{\partial}{\partial \beta} PPV &= -\tfrac{B(1-B)\alpha}{(B(1-\beta)+(1-B)\alpha)^2} \approx -\tfrac{\alpha B}{(\alpha+(1-\beta)B)^2}
\end{aligned}$$

Equations below show the derivatives of NPV.

$$\begin{aligned}
\tfrac{\partial}{\partial B} NPV &= -\tfrac{(1-\alpha)\beta}{((1-\alpha)(1-B)+B\beta)^2} \approx -\beta \\
\tfrac{\partial}{\partial \alpha} NPV &= -\tfrac{B(1-B)\beta}{((1-\alpha)(1-B)+B\beta)^2} \approx -B\beta \\
\tfrac{\partial}{\partial \beta} NPV &= -\tfrac{B(1-B)(1-\alpha)}{((1-\alpha)(1-B)+B\beta)^2} \approx -B
\end{aligned}$$

Equations below show the derivatives of Entropy of Intrusion Detector.

$$\frac{\partial}{\partial B} H(X) = \log\frac{1-B}{B} \approx -\log B\ ;\ \frac{\partial}{\partial\alpha} H(X) = 0\ ;\ \frac{\partial}{\partial\beta} H(X) = 0$$

Equations below show the derivatives of C_{ID}.

$$\begin{aligned}
\tfrac{\partial}{\partial B} C_{ID} &= \tfrac{\partial}{\partial B}(\tfrac{H(X)-H(X|Y)}{H(X)}) = \tfrac{1}{H(X)}\tfrac{\partial}{\partial B}(H(X|Y) - \tfrac{H(X|Y)}{H(X)^2}\tfrac{\partial}{\partial B}H(X) \\
&= \tfrac{1}{H(X)^2}(\log B(\alpha\log\tfrac{\alpha(1-B)}{B(1-\beta)+(1-B)\alpha} + (1-\alpha)\log\tfrac{(1-\alpha)(1-B)}{(1-\alpha)(1-B)+B\beta}) \\
&\quad -\log(1-B)((1-\beta)\log\tfrac{B(1-\beta)}{B(1-\beta)+(1-B)\alpha} + \beta\log\tfrac{B\beta}{(1-\alpha)(1-B)+B\beta})) \\
&\approx \tfrac{1}{H(X)^2}\left\{\alpha(\log\tfrac{\alpha}{\alpha+(1-\beta)B} - B\beta)\log B + B(\log\tfrac{(1-\beta)B}{\alpha+(1-\beta)B} - \beta\log\tfrac{1-\beta}{\beta(\alpha+(1-\beta)B)})\right\}
\end{aligned}$$

$$\begin{aligned}
\tfrac{\partial}{\partial \alpha} C_{ID} &= \tfrac{\partial}{\partial \alpha}(\tfrac{H(X)-H(X|Y)}{H(X)}) = -\tfrac{1}{H(X)}\tfrac{\partial}{\partial \alpha}(H(X|Y) \\
&= \tfrac{1-B}{H(X)}(\log\tfrac{\alpha(1-B)}{B(1-\beta)+(1-B)\alpha} - \log\tfrac{(1-\alpha)(1-B)}{(1-\alpha)(1-B)+B\beta}) \\
&\approx \tfrac{1}{H(X)}\log\tfrac{\alpha}{\alpha+(1-\beta)B}
\end{aligned}$$

$$\begin{aligned}
\tfrac{\partial}{\partial \beta} C_{ID} &= \tfrac{\partial}{\partial \beta}(\tfrac{H(X)-H(X|Y)}{H(X)}) = -\tfrac{1}{H(X)}\tfrac{\partial}{\partial \alpha}(H(X|Y) \\
&= -\tfrac{B}{H(X)}(\log\tfrac{B(1-\beta)}{B(1-\beta)+(1-B)\alpha} - \log\tfrac{B\beta}{(1-\alpha)(1-B)+B\beta}) \\
&\approx -\tfrac{B}{H(X)}\log\tfrac{1-\beta}{\beta(\alpha+(1-\beta)B)}
\end{aligned}$$

A.3 Approximation of C_{ID} for very small α and very large β

Substituting $1-\alpha \approx 1$ and $1-B \approx 1$ in equation 3 in section 2, we will get

$$\begin{aligned} H(X|Y) &= -B(1-\beta)\log\tfrac{B(1-\beta)}{B(1-\beta)+\alpha} - B\beta\log B\beta - (1-B)\log(1-B\beta) - \alpha\log\tfrac{\alpha}{B(1-\beta)+\alpha} \\ &= \underbrace{(-B(1-\beta)\log\frac{B(1-\beta)}{B(1-\beta)+\alpha} - \alpha\log\frac{\alpha}{B(1-\beta)+\alpha})}_{T1} + \underbrace{(-B\beta\log B\beta - (1-B)\log(1-B\beta))}_{T2} \end{aligned} \tag{11}$$

After substituting $\beta = 1$, second term of the above equation will become

$$T2 = -B\log B - (1-B)\log(1-B) = H(X)$$

For very small values of α, we can approximate $(1-exp(K\alpha^{1.19})) \approx K\alpha$. Thus the ROC curve (Section 4.1) becomes a straight line $\alpha = m(1-\beta)$ passing through the center. By substituting this in the first term,

$$\begin{aligned} T1 &= -B(1-\beta)\log\frac{B(1-\beta)}{B(1-\beta)+m(1-\beta)} - (m(1-\beta))\log\frac{m(1-\beta)}{B(1-\beta)+m(1-\beta)} \\ &= -(1-\beta)(B\log\frac{B}{B+m} + m\log\frac{m}{B+m}) \end{aligned}$$

As B and m are constant, we can write $K = B\log\frac{B}{B+m} + m\log\frac{m}{B+m}$, where K is a constant. Thus,

$$T1 = -K(1-\beta)$$

$$H(X|Y) = T1 + T2 = H(X) - K(1-\beta)$$

$$\begin{aligned} C_{ID} &= \frac{H(X)-H(X|Y)}{H(X)} = \frac{H(X)-H(X)+K(1-\beta)}{H(X)} \\ &= (1-\beta)\frac{K}{H(X)} \end{aligned}$$