# B-ROC Curves for the Assessment of Classifiers over Imbalanced Data Sets

**Alvaro A. Cárdenas** and **John S. Baras**
Department of Electrical and Computer Engineering
and The Institute for Systems Research
University of Maryland, College Park
{acardena,baras}@isr.umd.edu

**Abstract**

The class imbalance problem appears to be ubiquitous to a large portion of the machine learning and data mining communities. One of the key questions in this setting is how to evaluate the learning algorithms in the case of class imbalances. In this paper we introduce the Bayesian Receiver Operating Characteristic (B-ROC) curves, as a set of tradeoff curves that combine in an intuitive way, the variables that are more relevant to the evaluation of classifiers over imbalanced data sets. This presentation is based on section 4 of (Cárdenas, Baras, & Seamon 2006).

## Introduction

The term *class imbalance* refers to the case when in a classification task, there are many more instances of some classes than others. The *problem* is that under this setting, classifiers in general perform poorly because they tend to concentrate on the large classes and disregard the ones with few examples.

Given that this problem is prevalent in a wide range of practical classification problems, there has been recent interest in trying to design and evaluate classifiers faced with imbalanced data sets (Japkowicz 2000; Chawla, Japkowicz, & Kołcz 2003; Chawla, Japkowicz, & Kołz 2004).

A number of approaches on how to address these issues have been proposed in the literature. Ideas such as data sampling methods, one-class learning (i.e. recognition-based learning), and feature selection algorithms, appear to be the most active research directions for learning classifiers. On the other hand the issue of how to evaluate *binary* classifiers in the case of class imbalances appears to be dominated by the use of ROC curves (Ferri *et al.* 2004; 2005) (and to a lesser extent, by error curves (Drummond & Holte 2001)).

The class imbalance problem is of particular importance in intrusion detection systems (IDSs). In this paper we present and expand some of the ideas introduced in our research for the evaluation of IDSs (Cárdenas, Baras, & Seamon 2006). In particular we claim that for heavily imbalanced data sets, ROC curves cannot provide the necessary intuition for the choice of the operational point of the classifier and therefore we introduce the Bayesian-ROCs (B-ROCs). Furthermore we demonstrate how B-ROCs can deal with the uncertainty of class distributions by displaying the performance of the classifier under different conditions. Finally, we also show how B-ROCs can be used for comparing classifiers without any assumptions of misclassification costs.



## Performance Tradeoffs

Before we present our formulation we need to introduce some notation and definitions. Assume that the input to the classifier is a feature-vector $\mathbf{x}$. Let $C$ be an indicator random variable denoting whether $\mathbf{x}$ belongs to class zero: $C = 0$ (the majority class) or class one: $C = 1$ (the minority class). The output of the classifier is denoted by $A = 1$ if the classifier assigns $\mathbf{x}$ to class one, and $A = 0$ if the classifier assigns $\mathbf{x}$ to class zero. Finally, the class imbalance problem is quantified by the probability of a positive example $p = \Pr[C = 1]$.

Most classifiers subject to the class imbalance problem are evaluated with the help of ROC curves. ROC curves are a tool to visualize the tradeoff between the *probability of false alarm* $P_{FA} \equiv \Pr[A = 1|C = 0]$ and the *probability of detection* $P_D \equiv \Pr[A = 1|C = 1]$.

Of interest to us in the intrusion detection community, is that classifiers with ROC curves achieving traditionally "good" operating points such as $(P_{FA} = 0.01, P_D = 1)$ would still generate a huge amount of false alarms in realistic scenarios. This effect is due in part to the class imbalance problem, since one of the causes for the large amount of false alarms that IDSs generate, is the enormous difference between the large amount of normal activity compared to the small amount of intrusion events. The reasoning is that because the likelihood of an attack is very small, even if an IDS fires an alarm, the likelihood of having an intrusion remains relatively small. That is, when we compute the posterior probability of intrusion given that the IDS fired an alarm, (a quantity known as the *Bayesian detection rate*, or the *positive predictive value* (PPV)), we obtain:

$$PPV \equiv \Pr[C = 1|A = 1] = \frac{pP_D}{pP_D + (1-p)P_{FA}} \qquad (1)$$

Therefore, if the rate of incidence of an attack is very small, for example on average only 1 out of $10^5$ events is an attack

($p = 10^{-5}$), and if our detector has a probability of detection of one ($P_D = 1$) and a false alarm rate of 0.01 ($P_{FA} = 0.01$), then $\Pr[C = 1|A = 1] = 0.000999$. And so on average, of 1000 alarms, only one would be a real intrusion.

It is easy to demonstrate that the PPV value is maximized when the false alarm rate of our detector goes to zero, even if the detection rate also tends to zero! Therefore as mentioned in (Axelsson 1999) we require a trade-off between the PPV value and the *negative predictive value* (NPV):

$$\text{NPV} \equiv \Pr[C = 0|A = 0] = \frac{(1-p)(1-P_{FA})}{p(1-P_D)+(1-p)(1-P_{FA})} \quad (2)$$

However in the next section we point out that a tradeoff between $P_{FA}$ and $P_D$ (as in the ROC curves) as well as a tradeoff between PPV and NPV can be misleading for cases where $p$ is very small. That is, very small changes in the $P_{FA}$ and NPV values for our points of interest will have drastic performance effects on the $P_D$ and the PPV values.

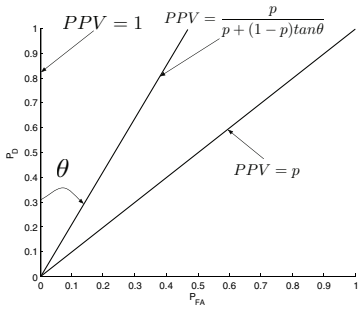


Figure 1: The PPV isolines in the ROC space are straight lines that depend only on θ. The PPV values of interest range from 1 to $p$

## B-ROC Curves

In order to make the intuition for the tradeoff between the PPV and the NPV clear, we now study the PPV and NPV isolines over the ROC curve space.

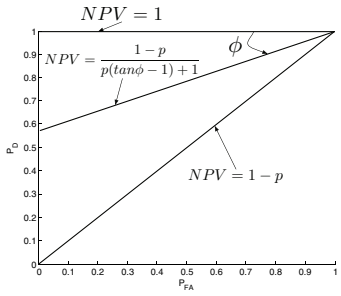


Figure 2: The NPV isolines in the ROC space are straight lines that depend only on ϕ. The NPV values of interest range from 1 to $1 - p$

**Fact 1** *Two sets of points* $(P_{FA1}, P_{D1})$ *and* $(P_{FA2}, P_{D2})$ *have the same PPV value if and only if*

$$\frac{P_{FA2}}{P_{D2}} = \frac{P_{FA1}}{P_{D1}} = \tan\theta \quad (3)$$

*where* θ *is the angle between the line* $P_{FA} = 0$ *and the isoline. Moreover the PPV value of an isoline at angle* θ *is*

$$PPV_{\theta,p} = \frac{p}{p+(1-p)\tan\theta} \quad (4)$$

*Similarly, two set of points* $(P_{FA1}, P_{D1})$ *and* $(P_{FA2}, P_{D2})$ *have the same NPV value if and only if*

$$\frac{1-P_{D1}}{1-P_{FA1}} = \frac{1-P_{D2}}{1-P_{FA2}} = \tan\phi \quad (5)$$

*where* ϕ *is the angle between the line* $P_D = 1$ *and the isoline. Moreover the NPV value of an isoline at angle* ϕ *is*

$$NPV_{\phi,p} = \frac{1-p}{p(\tan\phi - 1)+1} \quad (6)$$

Figures 1 and 2 show the graphical interpretation of Lemma 1. It is important to note the range of the PPV and NPV values as a function of their angles. In particular notice that as θ goes from 0° to 45° (the range of interest), the value of PPV changes from 1 to $p$. We can also see from Figure 2 that as ϕ ranges from 0° to 45°, the NPV value changes from one to $1 - p$. If $p$ is very small, then $NPV \approx 1$.

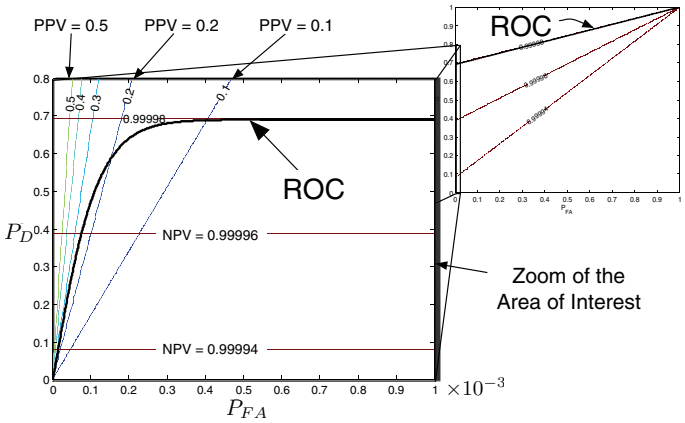


Figure 3: PPV and NPV isolines for the ROC of an IDS with $p = 6.52 \times 10^{-5}$

Figure 3 shows the application of Fact 1 to a typical ROC curve of an IDS. In this figure we can see the tradeoff of four variables of interest: $P_{FA}$, $P_D$, $PPV$, and $NPV$. Notice that if we choose the optimal operating point based on $P_{FA}$ and $P_D$, as in the typical ROC analysis, we might obtain misleading results because we do not know how to interpret intuitively very low false alarm rates, e.g. is $P_{FA} = 10^{-3}$ much better than $P_{FA} = 5 \times 10^{-3}$? The same reasoning applies to the study of PPV vs. NPV as we cannot interpret precisely small variations in NPV values, e.g. is $NPV = 0.9998$ much better

than $NPV = 0.99975$? Therefore we conclude that the most relevant metrics to use for a tradeoff in the performance of a classifier are $P_D$ and PPV, since they have an easily understandable range of interest.

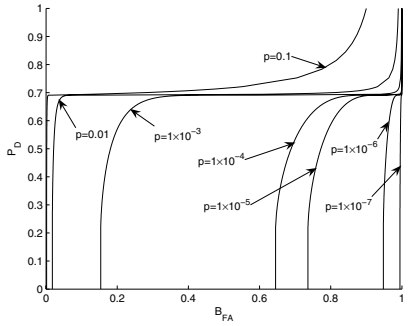


Figure 4: B-ROC for the ROC of Figure 3.

However, even when you select as tradeoff parameters the PPV and $P_D$ values, the isoline analysis shown in Figure 3 has still one deficiency, and it is the fact that there is no efficient way to account for the uncertainty of $p$. In order to solve this problem we introduce the B-ROC as a graph that shows how the two variables of interest: $P_D$ and $PPV$ are related under different severity of class imbalances. In order to follow the intuition of the ROC curves, instead of using $PPV$ for the x-axis we prefer to use 1-$PPV$. We use this quantity because it can be interpreted as the *Bayesian false alarm rate*: $B_{FA} \equiv \Pr[C = 0|A = 1]$. For example, for IDSs $B_{FA}$ can be a measure of how likely it is, that the operators of the detection system will loose their time each time they respond to an alarm. Figure 4 shows the B-ROC for the ROC presented in Figure 3. Notice also how the values of interest for the x-axis have changed from $[0, 10^{-3}]$ to $[0, 1]$.

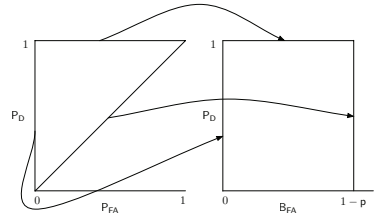


Figure 5: Mapping of ROC to B-ROC

In order to be able to interpret the B-ROC curves, Figure 5 shows how the ROC points map to points in the B-ROC. The vertical line defined by $0 < P_D \leq 1$ and $P_{FA} = 0$ in the ROC maps exactly to the same vertical line $0 < P_D \leq 1$ and $B_{FA} = 0$ in the B-ROC. Similarly, the top horizontal line $0 \leq P_{FA} \leq 1$ and $P_D = 1$ maps to the line $0 \leq B_{FA} \leq 1 - p$ and $P_D = 1$. A classifier that performs random guessing is represented in the ROC as the diagonal line $P_D = P_{FA}$, and this random guessing classifier maps to the vertical line defined by $B_{FA} = 1 - p$ and $P_D > 0$ in the B-ROC. Finally, to understand where the point $(0,0)$ in the ROC maps into the B-ROC, let $\alpha$ and $f(\alpha)$ denote $P_{FA}$ and the corresponding $P_D$ in the ROC curve. Then, as the false alarm rate $\alpha$ tends to zero (from the right), the Bayesian false alarm rate tends to a value that depends on $p$ and the slope of the ROC close to the point $(0,0)$. More specifically, if we let $f'(0^+) = \lim_{\alpha \to 0^+} f'(\alpha)$, then:

$$\lim_{\alpha \to 0^+} B_{FA} = \lim_{\alpha \to 0^+} \frac{\alpha(1-p)}{pf(\alpha) + \alpha(1-p)} = \frac{1-p}{p(f'(0^+) - 1) + 1}$$

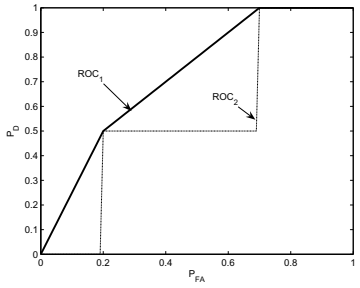


Figure 6: An empirical ROC ($ROC_2$) and its convex hull ($ROC_1$)

It is also important to recall that a necessary condition for a classifier to be optimal, is that its ROC curve should be concave. In fact, given any non-concave ROC, by following Neyman-Pearson theory, you can always get a concave ROC curve by randomizing decisions between optimal points (Poor 1988). This idea has been recently popularized in the machine learning community by the notion of the ROC convex hull (Provost & Fawcett 2001).

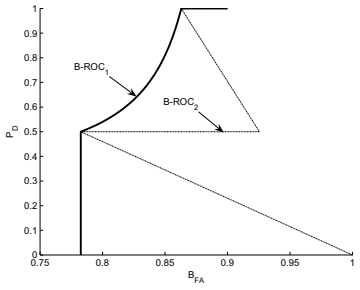


Figure 7: The B-ROC of the concave ROC is easier to interpret

The importance of this observation is that in order to guarantee that the B-ROC is a well defined continuous and non-decreasing function, we map only concave ROC curves to B-ROCs. In Figures 6 and 7 we show the only example in this paper of the type of B-ROC curve that you can get when you do not consider a concave ROC.

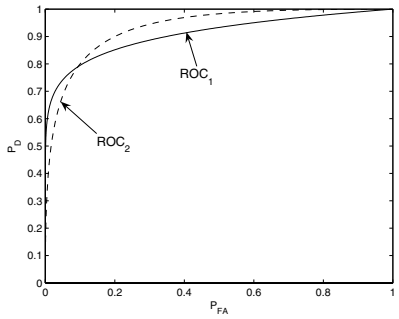


Figure 8: Comparison of two classifiers

We also point out the fact that the B-ROC curves can be very useful for the comparison of classifiers. A typical comparison problem by using ROCs is shown in Figure 8. Several ideas have been proposed in order to solve this comparison problem. For example by using decision theory as in (Cárdenas, Baras, & Seamon 2006; Provost & Fawcett 2001) we can find the optimal classifier between the two by assuming a given prior $p$ and given misclassification costs. However, a big problem with this approach is that the misclassification costs are sometimes uncertain and difficult to estimate a priori. With a B-ROC on the other hand, you can get a better comparison of two classifiers without the assumption of any misclassification costs, as can be seen in Figure 9.

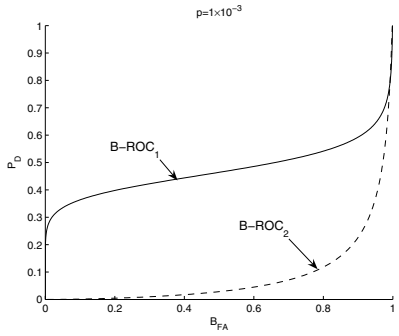


Figure 9: B-ROCs comparison for the $p$ of interest

## Conclusions

We believe that the B-ROC provides a better way to evaluate and compare classifiers in the case of class imbalances or uncertain values of $p$. First, for selecting operating points in heavily imbalanced environments, B-ROCs use tradeoff parameters that are easier to understand than the variables considered in ROC curves (they provide better intuition for the performance of the classifier). Second, since the exact class distribution $p$ might not be known a priori, or accurately enough, the B-ROC allows the plot of different curves for the range of interest of $p$. Finally, when comparing two classifiers, there are cases in which by using the B-ROC, we do not need cost values in order to decide which classifier would be better for given values of $p$. Note also that B-ROCs consider parameters that are directly related to exact quantities that the operator of a classifier can measure. In contrast, the exact interpretation of the expected cost of a classifier is more difficult to relate to the real performance of the classifier (the costs depend in many other unknown factors).

The work that is closest to ours is the analysis of information retrieval algorithms by using precision and recall. B-ROCs are an improvement over these metrics because they first uncouple the tradeoff curve from the class distribution in a given data set by means of first constructing the ROC of the classifier and then mapping it to different B-ROC curves. This makes B-ROCs less dependent on the testing data set, allows the evaluation of a classifier under uncertain $p$ values and facilitates any theoretical analysis of the B-ROC curves.

## Acknowledgements

This material is based upon work supported by the U.S. Army Research Office under Award No. DAAD19-01-1-0494 to the University of Maryland at College Park.

## References

Axelsson, S. 1999. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *Proceedings of the 6th ACM Conference on Computer and Communications Security (CCS '99)*, 1–7.

Cárdenas, A. A.; Baras, J. S.; and Seamon, K. 2006. A framework for the evaluation of intrusion detection systems. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy*.

Chawla, N. V.; Japkowicz, N.; and Kołcz, A., eds. 2003. *Proceedings of the International Conference for Machine Learning Workshop on Learning from Imbalanced Data Sets*.

Chawla, N. V.; Japkowicz, N.; and Kołz, A. 2004. Editorial: Special issue on learning from imbalanced data sets. *Sigkdd Explorations Newsletter* 6(1):1–6.

Drummond, C., and Holte, R. 2001. Explicitly representing expected cost: An alternative to ROC representation. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 198–207.

Ferri, C.; Flach, P.; Hernández-Orallo, J.; and Lachinche, N., eds. 2004. *First Workshop on ROC Analysis in AI*.

Ferri, C.; Lachinche, N.; Macskassy, S. A.; and Rakotomamonjy, A., eds. 2005. *Second Workshop on ROC Analysis in ML*.

Japkowicz, N., ed. 2000. *Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets*.

Poor, H. V. 1988. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 2nd edition.

Provost, F., and Fawcett, T. 2001. Robust classification for imprecise environments. *Machine Learning* 42(3):203–231.