# Modifying ROC Curves to Incorporate Predicted Probabilities

**Cèsar Ferri** CFERRI@DSIC.UPV.ES
DSIC, Universitat Politècnica de València
**Peter Flach** PETER.FLACH@BRISTOL.AC.UK
Department of Computer Science, University of Bristol
**José Hernández-Orallo** JORALLO@DSIC.UPV.ES
DSIC, Universitat Politècnica de València
**Athmane Senad** ASENAD@DSIC.UPV.ES
DSIC, Universitat Politècnica de València

## Abstract

The area under the ROC curve (AUC) is becoming a popular measure for the evaluation of classifiers, even more than other more classical measures, such as error/accuracy, logloss/entropy or precision. The AUC measure is specifically adequate to evaluate in two-class problems how well a model ranks a set of examples according to the probability assigned to the positive class. One shortcoming of AUC is that it ignores the probability values, and it only takes the order into account. On the other hand, logloss or MSE are alternative measures, but they only consider how well the probabilities are calibrated, and not its order. In this paper we introduce a new probabilistic version of AUC, called pAUC. This measure evaluates ranking performance, but also takes the magnitude of the probabilities into account. Secondly, we present a method for visualising a pROC curve such that the area under this curve corresponds to pAUC.

## 1. Introduction

Receiver Operating Characteristic (ROC) analysis (Provost & Fawcett, 1998; Swets, Dawes & Monahan, 2000; Provost & Fawcett, 2001) has been proven very useful for evaluating given classifiers in cases when the cost matrix or the final class distribution is not known when the classifiers were constructed. ROC analysis provides tools to select a set of classifiers that would behave optimally and reject sub-optimal classifiers. In order to do this, the convex hull of all the classifiers is constructed, giving a "curve" (a convex polygon). Additionally, the Area Under the ROC Curve, AUC, can be used as a simple measure that can be used to estimate how well a classifier acts as a ranker, i.e., a classifier that sorts the examples according to their probability of being positive.

ROC analysis and the AUC measure have been used extensively in the area of medical decision making (McClish, 1987; Hanley & McNeil, 1982; Mossman & Somoza, 1991; Zweig & Campbell, 1993), in the field of knowledge discovery, data mining, pattern recognition (Adams, & Hand, 1999) and science in general (Swets, Dawes & Monahan, 2000).

There are various ways to construct a ROC curve and calculate AUC. One of them is based in sorting the examples on their predicted probabilities of being positive, and accumulate the percentage of positives and negatives appearing from left to right (TPR and FPR). For instance, given five examples, with the following probabilities and classes: ($0.9^{+}$, $0.6^{-}$, $0.55^{+}$, $0.2^{-}$, $0.1^{-}$). We just sort the probabilities and indicate with a + or – if they are actual positives or negatives. In Figure 1 we show how to draw the ROC curve from the previous values. We just place them on a single axis (going from 1 to 0, as shown, or alternative, going from 0 to 1 if we use complementary probabilities). Then, going from left to right we move a segment up or right in the ROC space. The size of the segment is proportional to the number of positives or negatives, respectively. From this ROC curve we can compute the AUC. The following method for calculating AUC is from (Wu and Flach, 2005): for each positive on the sorted list, accumulate how many negatives follow it. Alternatively, we can accumulate for each negative how many positives precede it. For instance, the AUC of the above example is 5/6= 0.833, as shown in Figure 1.

Appearing in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005. 

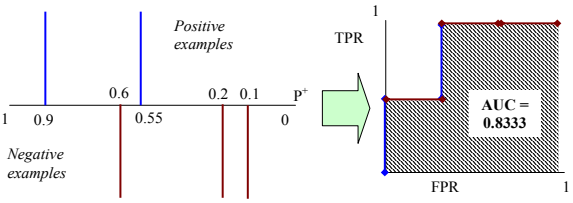



*Figure 1*. ROC curve and AUC based on ranking.

The problem of AUC is that it ignores the probability values, because it only takes the order into account. For instance, consider the following case: ($0.501^+$, $0.5^-$). The model does not clearly distinguish between the positive and the negative examples (their probabilities are almost the same), but the AUC is 1. A small variation of the given probabilities could give the following situation: (0.5-, 0.499+), where the area is 0, as seen in Figure 2.

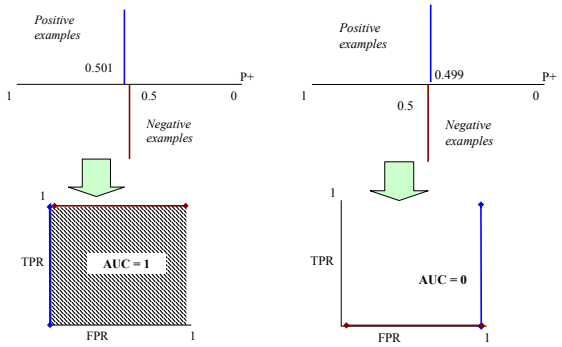



*Figure 2*. Small changes in probabilities can give rise to large changes in AUC.

One possible solution to this problem is the use of confidence intervals for ROC curves (Macskassy & Provost, 2004). However, it is not clear how these confidence bands can affect the AUC value when the probabilities are non-uniformly distributed from 1 to 0. Another option is to use alternative measures, such as MSE or RMS (Squared Error), logloss or cross-entropy or Kullback-Leibler Distance; for many authors, this is a better grounded measure for evaluating the quality of estimated probabilities. The MSE and logloss are alternative measures, but they only consider how well the probabilities are calibrated, and not to its order. The Brier score considers both separability and calibration. However, all these measures ignore ranking performance.

The outline of the paper is as follows. In Section 2 we define a probabilistic version of AUC, pAUC, and discuss how to draw pROC curves whose area is pAUC. This relies on a smoothing parameter *d*, and in Section 3 we discuss ways to compute *d*. Section 4 concludes.

This paper has similar aims to (Wu and Flach, 2005), the approaches are however different and have been developed independently.

## 2. pAUC and pROC curves

It is well-known that AUC is equivalent to the Wilcoxon sum of ranks test, which estimates the probability that a randomly chosen positive is ranked before a randomly chosen negative. The idea is to adapt this to obtain an estimate of the score difference between a randomly chosen pair of positive and negative examples. An unbiased estimator is obtained by the difference between the average scores of positives and the average scores of negatives. As this is a quantity between -1 and +1, it is actually a probabilistic version of the Gini coefficient, which is equal to the area below the curve above the diagonal. We can transform this into an AUC-like measure by adding 1 and dividing by 2. This gives the following definitions:

$$\text{pGINI} = \frac{\sum_{x\in\oplus} P(x)}{Pos} - \frac{\sum_{y\in\ominus} P(y)}{Neg}$$

$$\text{pAUC} = \frac{\frac{\sum_{x\in\oplus} P(x)}{Pos} - \frac{\sum_{y\in\ominus} P(y)}{Neg} + 1}{2} = \frac{\frac{\sum_{x\in\oplus} P(x)}{Pos} + \frac{\sum_{y\in\ominus} 1-P(y)}{Neg}}{2}$$

Here, $P(x)$ and $P(y)$ are the predicted probability of positive instance $x$ and negative instance $y$ to belong to the positive class, as before, and *Pos* and *Neg* are the total numbers of positive and negative examples. Thus, pAUC can be interpreted as the mean of the average positive score assigned to positives and the average negative score (predicted probability to belong to the negative class) assigned to negatives.

There are many examples where AUC > pAUC but also cases (more rarely) where pAUC > AUC. For instance, ($0.65^+$, $0.55^+$, $0.45^-$, $0.35^-$) has AUC=1 and pAUC=0.6. And ($1^+$, $0.1^-$, $0^+$) has AUC=0.5 and pAUC=0.7.

The main question is whether pAUC has a representation in the form of a curve. It would be useful to be able to have a corresponding curve from which the user can choose the threshold, according to the skew (cost matrix and class distribution) at application time. For instance, according to Figure 2, it is apparently the same to choose threshold 0.99 than to choose 0.55 for any skew (slope) greater than 1 (45°).

In this work we present a method for visualising a pROC curve, based on a representation of segments (with a uniform or normal distribution) of constant width or deviation, such that the area under this curve corresponds to pAUC. The key idea is to consider probabilities as intervals, i.e., to consider that a model giving probability 0.8 to an example means that the probability is "around 0.8". The simplest approximation to this idea is to consider segments of a fixed width *d*. For instance, Figure 3 illustrates how a curve and an area can be obtained with

this idea from the data: ($0.9^+$, $0.6^-$, $0.55^+$, $0,2^-$, $0.1^-$). As we can see on the right curve, there is a diagonal segment which corresponds to an area (around part of the second and third point) where we have a mixture of positive and negative influence.

One awkward thing about this method is that some segments overrun the interval [0,1]. There is no intuitive meaning on negative probabilities or probabilities greater than 1. However, these segments just try to model their area of influence in a symmetrical way. Obviously, we could have avoided exceeding the interval [0,1] but this would have complicated the mathematical and also the graphical interpretation of the final curves. Additionally, other possibilities could be to use segments around $p/(1-p)$, going then from 0 to $\infty$, with proportionally adjusted segments, where we avoid the problem of exceeding the intervals, but the segments would be asymmetrical.

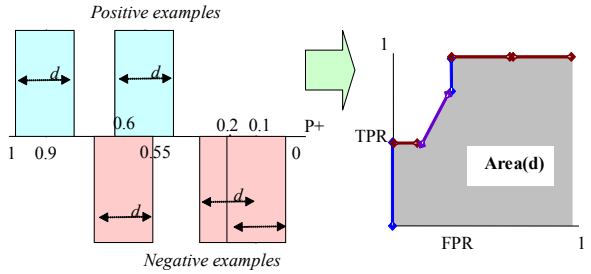



*Figure 3.* Drawing a pROC curve with probability segments.

With our approach, we just use the segments to show some kind of probability or smoothing reflecting the idea that two different test examples with closer probabilities could be wrongly ordered. The main question is which width $d$ to use. In order to analyse this, we have implemented a program that draws the curve and computes its area for different values of $d$. Figure 4 shows how the curve and the area evolves depending on the value of $d$ for the example $\{0.9^+, 0.8^-, 0.6^+, 0.3^-, 0.2^-\}$.

As we can see from the curves, the first curve ($d$=0) is logically equivalent to the original ROC curve. As long as $d$ increases, the area diminishes because the positive and negative influence regions overlap. The last curve ($d$=8) is almost the diagonal, with Area($d$)≈0.5. As we can see, $d$ can be greater than 1 and can exceed the interval [0,1]. Additionally, and more interestingly, there is a $d$ such that Area($d$) = pAUC, in particular, d= 1.596.

Summing up, in this case, it seems that:

$$\text{If } d = 0 \Rightarrow \text{Area}(d) = \text{AUC}$$
$$\text{If } d = 1.596 \Rightarrow \text{Area}(d) = \text{pAUC}$$
$$\text{If } d \to \infty \Rightarrow \text{Area}(d) \to 0.5$$

The results in the next section generalise this.

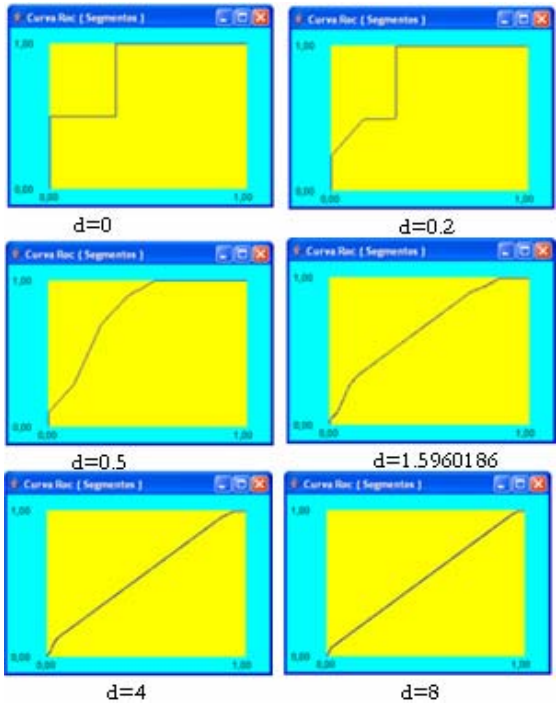



*Figure 4.* pROC curves for different values of $d$.

## 2.1 Some Theoretical Results

In what follows, we try to generalise the previous three cases.

THEOREM 1: For any set of examples labelled by probabilities and true labels, if $d = 0 \Rightarrow$ Area($d$) = AUC

Proof: Trivial, since originally the AUC is computed with each point with no width, as seen in Figure 1.

THEOREM 2: For any set of examples labelled by probabilities and true labels, if $d \to \infty \Rightarrow$ Area($d$) $\to 0.5$

Proof: If $d \to \infty$ then the overlapping between the segments is maximal or, in other words, the percentage of the regions with no overlapping is neglectable, i.e., 0. Consequently, we have that for every region in the density space we have that the same percentage of positives and negatives, and this means that the curve has always slope 0.5, given a diagonal from (0,0) to (1,1) in the ROC space, so making an Area = 0.5.

Before stating the last theorem (there exists a $d$ such that Area($d$) = pAUC), we have to show that Area($d$) is continuous,

PROPOSITION 3: Area($d$) is continuous, i. for all $\varepsilon>0$ then there exists a $\gamma>0$ such that for all $x$, $|x-d|<\gamma$ implies $|\text{Area}(x) - \text{Area}(d)| < \varepsilon$.

Proof: This is easy to see. If two examples have the same probability with opposite class, then with $d$=0, they are considered to overlap. If $d$ increases, the overlapping area

increases continuously with the increment of $d$. If two examples of different probability with opposite class, there is a $d=u$ ($d>0$) such that they start to overlap. This overlapping, at $d=u$ is 0, so it is continuous on the left. From that point the overlapping increases continuously from 0, so it is continuous on the right.

Nonetheless, it is important to see that Area($d$) is not necessarily increasing neither decreasing in all the domain. For instance, $\{1^+, 0.51^-, 0.49^+, 0^-\}$ has AUC= 0.75, pAUC=0.745. For $d=0$, as we have seen, Area($d$)=0.75. For d=$\infty$, Area($d$)=0.5, as we have seen as well. However, we can see that for $d=0.2$, Area($d$) = 0.874. Consequently, it increases initially but then decreases. In cases with more examples, this can have more local maxima and minima.

Now that we know that Area($d$) is continuous, we need to analyse the relationship with AUC.

First, we give definitions for AUC, pAUC and Area($d$) without using ranks:

$$AUC = \frac{1}{Pos \cdot Neg} \sum_{x \in \oplus} \sum_{y \in \Theta} g(x,y)$$

with

$$g(x,y) = \begin{cases} 1 & if \quad x > y \\ 0.5 & if \quad x = y \\ 0 & if \quad x < y \end{cases}$$

$$pAUC = 0.5 + \frac{1}{2} \frac{1}{Pos \cdot Neg} \sum_{x \in \oplus} \sum_{y \in \Theta} d(x,y)$$

with

$$d(x,y) = x - y$$

$$Area(d) = \frac{1}{Pos \cdot Neg} \sum_{x \in \oplus} \sum_{y \in \Theta} g_d(x,y)$$

with

$$g_d(x,y) = \begin{cases} g(x,y) & if \quad |x-y| \geq d \\ 1 - \frac{(1-c)^2}{2} \text{ with } c = \frac{|x-y|}{d} & if \quad |x-y| < d \text{ and } x > y \\ \frac{(1-c)^2}{2} \text{ with } c = \frac{|x-y|}{d} & if \quad |x-y| < d \text{ and } x \leq y \end{cases}$$

THEOREM 4: For any set of examples labelled by probabilities and true labels, there exists a real number $d \geq 0$ such that Area($d$) = pAUC.

Proof: We consider four mutually exclusive and exhaustive cases.

Case a) Let us consider that AUC ≥ pAUC and pAUC ≥ 0.5. Necessarily, since Area(0) = AUC and $\lim_{d \to \infty}$ Area($d$) = 0.5, and Area($d$) is continuous, there exists a $d \geq 0$ such that Area($d$) = pAUC.

Case b) Let us consider that AUC ≤ pAUC and pAUC ≤ 0.5. Necessarily, since Area(0) = AUC and $\lim_{d \to \infty}$ Area($d$) = 0.5, and Area($d$) is continuous, there exists a $d \geq 0$ such that Area($d$) = pAUC.

Case c) Let us consider that AUC ≥ pAUC and pAUC ≤ 0.5. Since pAUC ≤ 0.5, the term:

$$\frac{1}{Pos \cdot Neg} \sum_{x \in \oplus} \sum_{y \in \Theta} d(x,y)$$

must be negative. This implies that the average overlapping is negative. This clearly means that there are correctly ordered pairs (positives before negatives) which are closer than other inverted pairs (negatives before positives). Otherwise, pAUC couldn't be lower than 0.5 and lower than AUC. This means that with increasing values of d we can further decrease the Area(d), which at d=0 is equal to AUC. We can decrease this until pAUC (at least) since $g_d(x,y)$ grows quicker than $d(x,y)$.

Case d) Let us consider that AUC ≤ pAUC and pAUC ≥ 0.5. Since pAUC ≥ 0.5, the term:

$$\frac{1}{Pos \cdot Neg} \sum_{x \in \oplus} \sum_{y \in \Theta} d(x,y)$$

must be positive. This implies that the average overlapping is positive. This clearly means that there are inverted pairs (negatives before positives) which are closer than other correctly ordered pairs (positives before negatives). Otherwise, pAUC couldn't be greater than 0.5 and greater than AUC. This means that with increasing values of d we can further increase the Area(d), which at d=0 is equal to AUC. We can increase this until pAUC (at least) since $g_d(x,y)$ grows quicker than $d(x,y)$.

We have shown that there exists a $d$ such that Area($d$) = pAUC. However, is this $d$ unique? The answer is no as we can see in the following example.

**Example**: Consider the ranking $\{1^+, 1^+, 0.6^-, 0.6^-, 0.50001^+, 0.49999^-, 0.45^+, 0.45^+, 0^-, 0^-\}$. The AUC of this ranking is 0.68. The pAUC is approximately 0.67. Logically Area(0)= 0.68. For $d=0.05$, we have that Area($d$) = 0.66 approx. Consequently, there exists a d between 0 and 0.05 which makes Area($d$) = pAUC. But for $d=0.4$ we have that Area($d$) = 0.732. Consequently, from $d=0.05$ with Area($d$) = 0.66 and with $d=0.4$ with Area($d$) = 0.732, there must be an intermediate $d$ such that Area($d$) = pAUC. Since this $d$ is clearly different from the previous one, this problem shows two values of d such that Area($d$) = pAUC.

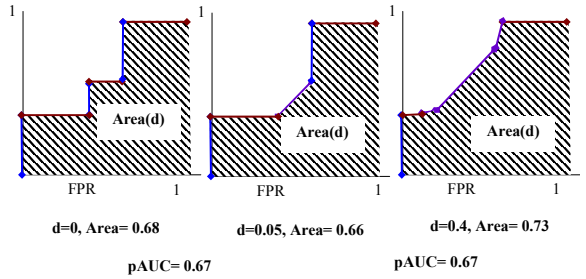



*Figure 5.* Oscillating area for increasing values of d gives two different d where Area(d) = pAUC.

The entire evolution of Area(*d*) is shown in Figure 6.

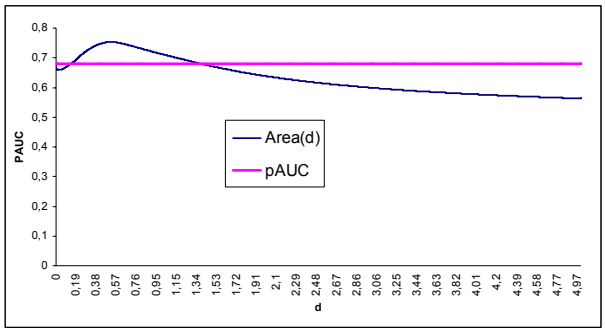



*Figure 6.* Oscillating area for increasing values of *d* gives two different d where Area(*d*) = pAUC.

### 2.2 Interpretation of *d*

After the previous results, let us stop for analysing the interpretation of *d*. pAUC can be seen as a smoothing of AUC. The use of segments Area(*d*) also acts as a smoothing, where the probability is distributed over an area (uniformly) instead to a single point (which is quite unlikely). This has a natural interpretation of Area(*d*). However, the value of *d* to be used is, however, would be a big question.

Theorem 4 ensures that we can always construct a pROC curve that "expresses" a given pAUC, which is a modified ROC curve. This curve is closer to the original ROC curve for small *d* and more different for big *d*. The pAUC value can be seen as a way to determine the *d* for the Area(*d*) and thus finding an adequate degree of smoothing. The Area(*d*), on the other hand, is useful to find a graphical interpretation to pAUC.

The values of *d* can be interpreted as points where both degrees of smoothing match. As we have seen, there can be more than one *d* with Area(*d*) = pAUC. Any of these *d* has a meaning, however, in what follows we choose the canonical *d* as the smallest one. This is justified because the smallest *d* is, the curve is closer to the original AUC curve.

### 2.3 Using normally distributed segments

Even though in the previous sections we have worked with uniform segments (rectangles), because of their simplicity, it seems that the interpretation of a probability such as 0.8, as a probability around 0.8 does not have its best interpretation with a uniform distribution with crisp intervals. A more natural option would be to consider a normal (Gaussian) distribution centered on the probability, where the factor to gauge is also called *d*, but represents the double of the standard deviation.

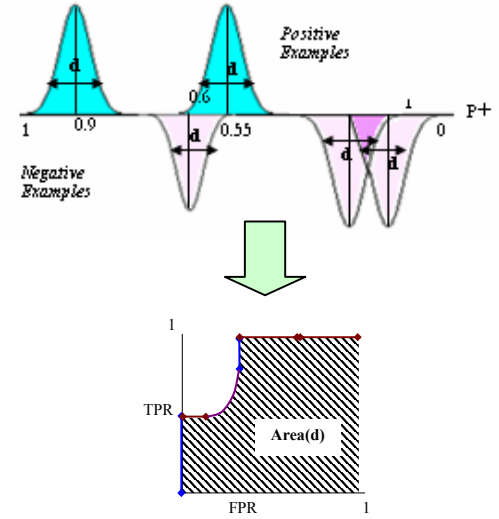



*Figure 7.* Using normally distributed segments to obtain a pROC curve. Here, *d* is shown to be twice the standard deviation.

Nonetheless, it is expected to have similar results (both theoretically and practically). Additionally some of the curves are expected to be "smoother".

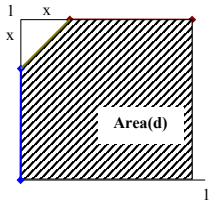
## 3. How "*d*" is calculated to have Area(*d*) = pAUC

It doesn't seem easy to find an analytical way to compute the minimum d such that Area(*d*) = pAUC. In the following subsection we sketch how it could be for a simple example.

### 3.1 Computing *d* analytically

In the last sections we have shown that there is at least one value for *d* where the value of the area is the pAUC. Can this value be found analytically? We have researched in this direction without positive results. The behaviour of the pROC curve is very dependent on the problem configuration (number and probabilities), and we believe that there is not a direct way to compute the value for *d* given a set of positive and negative examples.

Let us illustrate the complication of this computation with a very simple setting. Consider a positive example with value *a*, and a negative example *b*, where *a*>*b*. Clearly, AUC=1 and pAUC=(*a*–*b*+1)/2. The pROC curve will have the following shape:




If Area($d$)=pAUC, using the pROC curve, we have $x^2/2=1-\text{pAUC}=1-(a-b+1)/2$. From the segment representation we have $x=(b-a+d)/d$. Simplifying these expressions we get $d=\frac{a-b}{1-\sqrt{b-a+1}}$. For instance, if we have $a$=0.6, and $b$=0.4, then $d$=1.89. However, it is difficult to obtain direct formulas for more complicated cases.

### 3.2 Computing *d* numerically. Examples

Instead of an analytical solution, we propose a numerical solution. The proposed solution is an iterative method that finds the value $d$ numerically. It is based on increasing $d$ from 0 to $\infty$ in small $\varepsilon$ since the proper value is found. This is made from left to right to find the smallest $d$ possible such that Area($d$)=pAUC. The $\varepsilon$ is chosen sufficiently small to avoid passing through the correct $d$. Also, a small $\varepsilon$ also ensures that the difference between the pAUC and the Area($d$) is usually small (< 0.001).

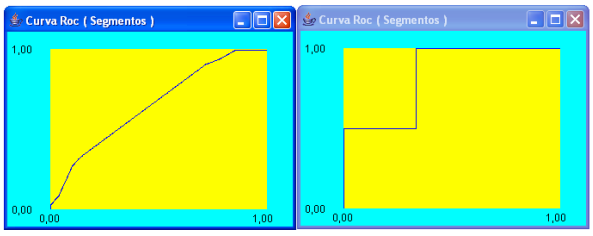
Below, we show some examples with the ROC curve and the corresponding pROC curve.

***Example 1***

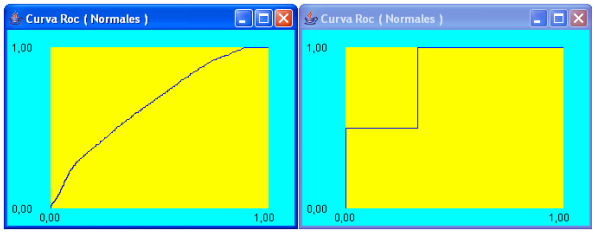
$\{ 0.9^+ \;\; 0.8^- \;\; 0.6^+ \;\; 0.3^- \;\; 0.2^- \}$

AUC: 0.833 pAUC: 0.658

Uniform pROC Curve for $d$ = 1.65 and for $d$=0:




Normal pROC Curve for $d$ = 1.74 and for $d$=0:



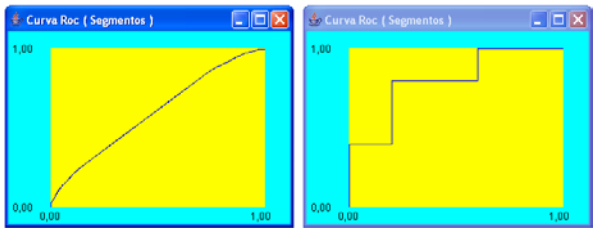

In this case, the smoothing is strong ($d$ is relatively high) and the pROC curve is significantly different to the ROC curve. Nonetheless, some similarities remain, such as the concavity at the left-bottom part of the smoothed curves, and the top-right segment where the true positive rate is 1.

***Example 2***

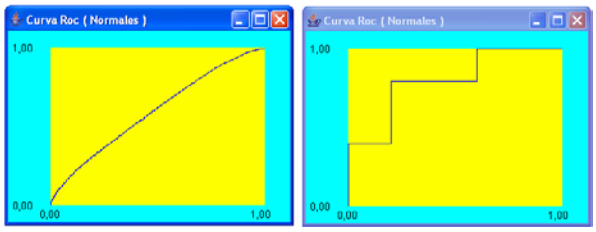
$\{0.85^+ \; 0.78^+ \; 0.7^- \; 0.55^+ \; 0.52^+ \; 0.5^- \; 0.4^- \; 0.3^+ \; 0.25^- \; 0.15^- \}$

AUC: 0.8 pAUC: 0.6

Uniform pROC Curve for $d$ = 1.71 and for $d$=0:




Normal pROC Curve for $d$ = 1.8 and for $d$=0:




In this case, we can see a more general situation with more points. We can see that the pROC curve acts as a kind of smoothing of the ROC curve.

### 3.3 General interpretation of the pROC curves

From the previous examples we can get an idea of what pROC curves mean.

- They smooth the ROC curves, especially when differences between probabilities of consecutive examples of different class are small.
- The value is found globally, so it can give high overlap in some areas but small overlap in others. So, especially when $d$ is not too high, some parts of the ROC curve might seem unaffected (this is especially the case because we use a uniform distribution and not normal distribution).
- The greater the number of examples, the difference of probabilities between consecutive examples of different classes usually becomes smaller and hence the pROC curve has a closer (but smoother) correspondence to the original ROC curve.

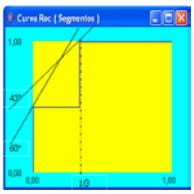
Summing up, as long as the number of example is greater both curves tend to be more similar, because both are softer. The differences are more important especially when the ROC curve is not very reliable: few examples and small differences in probabilities.

A second issue is the use of the new curves for

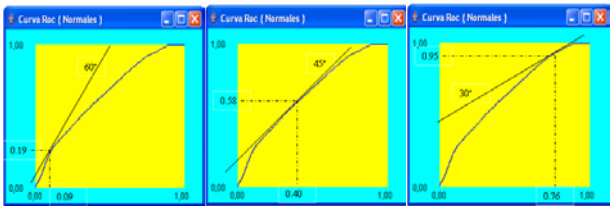
establishing probability thresholds according to some application skew (cost matrix + class distribution). In classical ROC analysis, given the skew at application time, we get the slope of an iso-accuracy line, which gives us the threshold (or the threshold interval) which would give the best accuracy for that skew.

With pROC curves, we can do the same thing, and, interestingly, we get different results. We will just reconsider example 1 seen before: { $0.9^+$ $0.8^-$ $0.6^+$ $0.3^-$ $0.2^-$ }, because we can see that differences are strong and also because we see some awkward results.

Using the original ROC curve, we see that slopes greater than 55° (approx.) cut the curve at point 0% FPR and 50% TPR. This means that we would choose a probability between 0.9 and 0.8 as a threshold. For slopes lower than 55° (approx.), the cutpoint is at 33% FPR and 100% TPR. This means that we would choose a probability threshold between 0.6 to 0.3.



On the other hand, if we use the pROC curves, we have a different picture:



The first picture shows the cut point with a slope of 60º. In this case, we have FPR=9% and TPR=19% (which is very different from FPR=0% and TPR=50% as in the ROC curve). Using the original probabilities, this would mean that the threshold should be between 0.9 and 0.8 as before. Similarly, we can compute these values for the other two skews, as shown in the following table:

| *Skew (slope)* | *FPR/TPR in ROC curve* | *Threshold in ROC curve* | *FPR/TPR in pROC curve* | *Threshold in pROC using the original prob.* |
|---|---|---|---|---|
| 60º | 0% / 50% | [0.9, 0.8] | 9% / 19% | [0.9, 0.8] |
| 45º | 33% / 100% | [0.6, 0.3] | 40% / 58% | [0.8, 0.6] |
| 30º | 33% / 100% | [0.6, 0.3] | 76% / 95% | [0.3, 0.2] |

From the previous table we see that the FPR/TPR are much more gradual than for the original ROC curve. This is usually the case when we have few examples. However, note that using smoothed probabilities (the segments or the Gaussian function) means that we can get values outside the interval [0, 1], which are difficult to interpret as a threshold.

In general, if we consider examples with more points, it is less likely to get thresholds outside [0, 1] and the use of pROC to establish the threshold could be more reliable. However, it would be necessary to perform an empirical analysis on this.

## 4. Conclusions

Several works have demonstrated that accuracy has many drawbacks as an evaluation measure for classification models. The Area Under the ROC curve (AUC) has been promoted as a more suitable alternative for evaluating classifiers. However, this measure is only concentrated in how the model ranks the examples, without taking into account the probabilities associated to the examples. We have illustrated this behaviour by means of a simple example where we illustrate how a slight variation of the confidences produces an abrupt change in the AUC.

Having in mind these limitations, a new measure, namely probabilistic Area Under the ROC curve (pAUC) has been defined. In this case both aspects (ranking and probabilities) are considered. Although pAUC avoids many of the drawbacks of the AUC measure, it apparently loses one of the strongest points of the AUC measure: the possibility of representing the performance of the evaluated model, i.e. the ROC curves.

In this work, we introduce a representation of the pROC curves. The basic idea of drawing pROC curves is based on the original ROC curves, but in this case we consider probabilities as intervals. These segments are set to a width *d*, and varying this value we are able to find a curve whose area is exactly the pAUC of the model. For constructing the intervals we have studied two different approaches: uniform or normal distribution. Both approaches obtain similar results. We have also presented some theoretical results of the proposed representation.

pROC curves help to know different details about the performance of the evaluated model with respect to classical ROC curves do. Although pROC curves have a similar shape to classical ROC curves, they are usually smoother. For a high number of examples, both curves tend to be more similar, because both are softer. In general, we can find the main differences when the ROC curve is not very reliable: few examples and small differences in probabilities. We have also used the curves to obtain cut points. However, obtaining a good probability threshold from here has shown to present many caveats. This suggests to study asymmetrical segments, where we never exceed the interval [0..1]. Another option would be a normalisation of the probabilities after applying the segments of size *d* or the Gaussian function.

Currently we are working on an experimental evaluation on whether pAUC is a good selection measure (when we do not know the skew), and also showing experimentally whether there is a method to establish the threshold from pROC curves (when we know the skew) that can be better

than the threshold obtained by ROC curves.

Finally, although the approach is different, we would like to connect this work with other previous works on confidence bands for ROC curves, as well as comparing to other measures, theoretically and experimentally.

## Acknowledgements

We would like to thank Shaomin Wu for some discussions around the idea of incorporating probabilities into the AUC measure. This work has been partially supported by the EU (FEDER) and the Spanish MEC, under grant TIN 2004-7943-C04-02 and Generalitat Valencian under grant GV04A-389.

## References

Adams, N.M. and Hand, D.J. (1999) "Comparing classifiers when the misallocation costs are uncertain" *Pattern Recognition*, Vol. 32 (7) (1999) pp. 1139-1147.

Hanley, J.A. and McNeil, B.J. (1982) "The meaning and use of the area under a receiver operating characteristic (ROC) curve" *Radiology*. 1982: 143:29-36.

McClish, D.K. (1987) "Comparing the areas under more than two independent ROC curves" *Med. Decis. Making*, 1987; 7:149-55.

Macskassy, S.A. and Provost, F.J. (2004) ·Confidence Bands for ROC Curves: Methods and an Empirical Study" ROCAI 2004: 61-70, 2004.

Mossman, D. and Somoza, E. (1991) "ROC curves, test accuracy, and the description of diagnostic tests" *J. Neuropsychiatr. Clin. Neurosci.* 1991, 3: 330-3.

Provost, F. and Fawcett, T. (1997) "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distribution" *Proc. of The Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pp. 43-48, Menlo Park, CA: AAAI Press.

Provost, F. and Fawcett, T. (2001) "Robust Classification for Imprecise Environments" *Machine Learning* 42, 203-231, 2001

Swets, J., Dawes, R., and Monahan, J. (2000) "Better decisions through science" *Scientific American*, October 2000, 82-87.

Wu, S. and Flach, P.A. (2005) "Scored and Weighted AUC Metrics for Classifier Evaluation and Selection", 2nd Workshop on ROC Analysis in Machine learning, ROCML'05.

Zweig, M.H. and Campbell, G. (1993) "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine", *Clin. Chem*, 1993; 39: 561-77.