
The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics

Peter A. Flach

Department of Computer Science, University of Bristol, Woodland Road, Bristol BS8 1UB, United Kingdom

PETER.FLACH@BRISTOL.AC.UK

Abstract

Many different metrics are used in machine learning and data mining to build and evaluate models. However, there is no general theory of machine learning metrics, that could answer questions such as: When we simultaneously want to optimise two criteria, how can or should they be traded off? Some metrics are inherently independent of class and misclassification cost distributions, while other are not — can this be made more precise? This paper provides a derivation of ROC space from first principles through 3D ROC space and the skew ratio, and redefines metrics in these dimensions. The paper demonstrates that the graphical depiction of machine learning metrics by means of ROC isometrics gives many useful insights into the characteristics of these metrics, and provides a foundation on which a theory of machine learning metrics can be built.

1. Introduction

Many different metrics are used in machine learning and data mining to build and evaluate models. For instance, for model building we can use precision in (association) rule learning, information gain in decision tree building, and weighted relative accuracy in subgroup discovery. For model evaluation on a test set we can use accuracy, F-measure, or area under ROC curve. Many variants of these metrics are used, for instance probability estimates can be Laplace-corrected or m -corrected to take a prior distribution into account.

However, which metrics are used in a certain context seems to be, to a large extent, historically determined. For instance, it is not clear why precision is a good metric for deciding which condition to add to a classification rule, or why decision tree splitting criteria should rather be impurity-based. Machine learning researchers have (re-)discovered the importance of being able to deal with skewed class and misclassification cost distributions that may differ from training to deployment, but a general theory how to characterise dependence on these aspects is lacking.

In this paper we use ROC analysis to start tackling these and related issues. Our main tool will be ROC isometric plots, which are contour plots for the metric under investigation. ROC isometrics are a very powerful tool to analyse and characterise the behaviour of a range of machine learning metrics. While ROC analysis is commonly associated with model selection, we demonstrate in this paper that ROC analysis has a much wider applicability and should be one of the most used instruments in every machine learner's toolbox.

There has been some previous work in this area. Provost and Fawcett (2001) use isometrics (which they call iso-performance lines) to determine the optimal point on a ROC convex hull. The term 'isometric' seems to originate from (Vilalta & Oblinger, 2000); they give a contour plot for information gain identical to the one presented in this paper, but their analysis is quantitative in nature and the connection to ROC analysis is not made. The purpose of this paper is to outline a general framework for analysing machine learning metrics, and to demonstrate the broad applicability of the framework, which ranges from classification, information retrieval, subgroup discovery, to decision tree splitting criteria. To demonstrate that the approach can lead to concrete, useful results, we derive an equivalent simplification of the F-measure used in information retrieval, as well as a version of the Gini splitting criterion that is insensitive to class or cost distributions. Further results are obtained in (Fürnkranz & Flach, 2003) by applying a similar analysis to rule evaluation metrics.

The outline of the paper is as follows. Section 2 presents the fundamentals of ROC space, including a novel perspective on how 2D ROC space is obtained from 3D ROC space by means of the skew ratio. Section 3, the main part of the paper, analyses a range of evaluation metrics and search heuristics through their isometric plots. Section 4 presents a more formal analysis, and Section 5 concludes.

2. 3D and 2D ROC Space

A *contingency table* (sometimes called *confusion matrix*) is a convenient way to tabulate statistics for evaluating the quality of a model. In Table 1, TP , FP , TN and FN stand for true/false positive/negative counts, respectively; PP and PN stand for predicted positive/negative; and POS

and *NEG* stand for actual positive/negative. N is the sample size. We use lowercase for relative frequencies, e.g. $tp = TP/N$ and $pos = POS/N$. In this paper we will only consider metrics that can be defined in terms of the counts in a contingency table (this excludes, e.g., metrics that consider model complexity). We also restrict attention to two-class problems.

Table 1. Counts organised in a two-by-two contingency table with marginals. The top two rows stand for actual classes, while the left two columns stand for predicted classes.

| | | |
|-----------|-----------|------------|
| <i>TP</i> | <i>FN</i> | <i>POS</i> |
| <i>FP</i> | <i>TN</i> | <i>NEG</i> |
| <i>PP</i> | <i>PN</i> | <i>N</i> |

The metrics we consider all have in common that they evaluate the quality of a contingency table in some way. However, it is not necessarily the case that the counts are obtained from evaluating a complete model; they could also be obtained from parts of a model, such as a single rule, or a decision tree split. So, even though ROC analysis is usually applied to the model evaluation stage rather than the model building stage, the analysis in this paper applies to any metric that evaluates the quality of a two-by-two contingency table for some purpose, and we use the term ‘model’ in a generic sense.

2.1 Contingency Tables in 3D

A two-by-two contingency table with relative frequencies has three degrees of freedom, and can thus be seen as a point in a three-dimensional Cartesian space. The chosen co-ordinates depend on the purpose of the analysis. A typical choice¹ related to ROC analysis is to use the *false positive rate* $fpr = FP/NEG$ on the X-axis, *true positive rate* $tpr = TP/POS$ on the Y-axis, and the relative frequency of positives $pos = POS/(POS+NEG)$ on the Z-axis. We will call this *3D ROC space*. As we explain in Section 2.2, the key assumption of ROC analysis is that true and false positive rates describe the performance of the model independently of the class distribution, and we are thus free to manipulate the Z-axis in 3D ROC space to conform to the class distribution of the environment in which the model will be employed.

Any metric that is defined in terms of the counts in a contingency table assigns a value to each point in 3D ROC space. For instance, accuracy can be defined as $pos * tpr + (1-pos) * (1-fpr)$. The set of points that are assigned the

same accuracy m are given by the equation $pos * tpr + (1-pos) * (1-fpr) = m$, which can be re-arranged to yield the surface $pos = (m + fpr - 1) / (tpr + fpr - 1)$ under the constraints $0 \leq pos \leq 1$ and $tpr + fpr - 1 \neq 0$. (These constraints are needed because not all combinations of tpr and fpr are possible for a given value of accuracy: e.g., on the line $tpr + fpr - 1 = 0$ we have equal true positive and true negative rates, hence $tpr = 1 - fpr = m$ is the only possible point on this line.) We call this surface an *accuracy isosurface*; a graphical depiction is given in Figure 1. The intuition behind this isosurface is that in the bottom plane $pos = 0$ we have the line $1 - fpr = m$; increasing pos rotates this line around $tpr = 1 - fpr = m$ until in the top plane we reach the line $tpr = m$. By increasing m the surface is shifted to the front and left, as indicated in Figure 1.

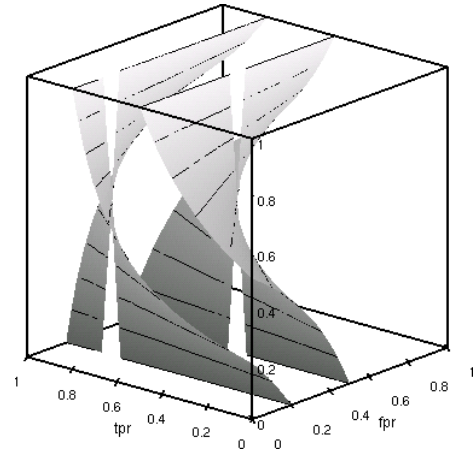


Figure 1. Accuracy isosurfaces for 80% (front/left) and 50% accuracy in 3D ROC space, with the proportion of positives on the vertical axis.

2.2 From 3D to 2D: Skew Ratio

Suppose we have evaluated a number of models and plotted their contingency tables in 3D ROC space. The models may have been evaluated on different test sets with different class distributions, hence these points may be located in different horizontal planes. Assuming that *the models’ true and false positive rates are independent of the class distribution in the test set*, we are free to use vertical projection to collect all points in a single plane. We could decide to discard the vertical dimension altogether and work in 2D (fpr, tpr) space. However, metrics such as accuracy have different behaviour in different ‘slices’ of ROC space (see Figure 1). It is therefore better, at least conceptually, to regard 2D ROC space as a horizontal slice of 3D ROC space upon which all 3D ROC points describing models are projected.

The appropriate slice can be selected once the expected class distribution in the target context is known, which fixes the behaviour of the metrics involved. To ease subsequent analysis, we will use the *class ratio* $c = NEG/POS$ rather than the relative frequency of positives; note that this is merely a re-scaling of the vertical axis in 3D ROC

¹ Other choices are possible; e.g., in information retrieval (Van Rijsbergen, 1979) performance is evaluated irrespective of the true negatives (non-answers that are correctly not returned) and the sample size, and the chosen co-ordinates are *precision* $= TP/PP$ and *recall* $= TP/POS$ (the same as tpr). DET curves (Martin *et al.*, 1997) are essentially a re-scaling of ROC curves.

space ($pos=1/(1+c)$). If all models were evaluated on a test set with class ratio equal to the expected class ratio, all points in 3D ROC space are in the same horizontal plane and we use this slice as our 2D projection. Otherwise, we project all points onto the appropriate slice corresponding to the expected class ratio.

It is easy to factor in non-uniform misclassification costs by adjusting the expected class ratio. For instance, if false positives are 3 times as expensive as false negatives, we multiply the expected class ratio with 3 — the intuition is that this cost ratio makes us work harder on the negatives. This can be further extended by taking correct classification profits into account. For instance, if true positives result in profit 5, false negatives result in cost 3, true negatives have profit 2, and false positives have cost 1, then this results in an adjustment to the class ratio of $(2-1)/(5-3)$: the cost matrix has effectively a single degree of freedom (Elkan, 2001).

All these scenarios (test class ratios are meaningful, test class ratios need adjustment, misclassification costs need to be taken into account, correct classification profits need to be taken into account as well) are thus covered by assuming a single *skew ratio* c : $c < 1$ tells us positives are more important, $c > 1$ tells us the opposite. It is therefore perfectly legitimate to assume in what follows the simplest scenario: all models are evaluated on the same test set with meaningful class distribution, and c stands for the ratio of negatives to positives in the test set. The reader should just keep in mind that the analysis is equally applicable to the other scenarios, in which the skew ratio only partly depends on the class distribution in the test set.

To summarise, we assume that true and false positive rates are sufficient statistics for characterising the performance of a classifier in any target context. The skew ratio tells us what the expected trade-off between negatives and positives is in the target context, and is therefore a parameter of the metrics we consider. If the skew ratio is biased (i.e., unequal to 1), it is irrelevant for our purposes whether this is because of a skewed class distribution, unequal misclassification costs, or both. We will therefore avoid terms like ‘cost-sensitive’ in favour of the more neutral *skew-sensitive*. Only if we want to interpret what a metric measures, we need to take the components of the skew ratio into account. For instance, consider accuracy: (1) Disregarding misclassification costs, accuracy estimates the probability that a randomly chosen example is correctly classified. (2) With misclassification costs, accuracy estimates the probability that a randomly chosen example incurs zero cost. (3) With misclassification costs and correct classification profits, accuracy estimates the probability that a randomly chosen example incurs a profit. Of course, if we want to know the expected *yield* of a model (the number of correctly classified examples, the amount of cost or profit incurred) we need to know, in addition, the absolute numbers of examples of each class and the associated cost and profit parameters.

2.3 2D ROC Space

2D ROC space, hereafter simply referred to *ROC space* if no confusion can arise, is thus a slice out of 3D ROC space determined by the skew ratio c . The skew ratio does not influence the position of models in the (fpr, tpr) plane, but it does influence the behaviour of metrics which take c as a parameter. Isosurfaces in 3D ROC space become lines in 2D ROC space, called *isometrics* in this paper. We will take a detailed view at ROC isometrics for a range of machine learning metrics in the next section. In the remainder of the present section, we will take a closer look at some specific points and lines in 2D ROC space.

The points (0,0) and (1,1) represent the training-free classifiers *AlwaysNegative* and *AlwaysPositive*; the point (0,1) represents the ideal classifier, and (1,0) represents the classifier which gets it all wrong. The ascending diagonal (0,0)–(1,1) represents random training-free behaviour: any point (p,p) can be obtained by predicting positive with probability p and negative with probability $(1-p)$. The upper left triangle contains classifiers that perform better than random, while the lower right triangle contains those performing worse than random. The descending diagonal (0,1)–(1,0) represents classifiers that perform equally well on both classes ($tpr=1-fpr=tnr$); left of this line we find classifiers that perform better on the negatives than the positives, to the right performance on the positives dominates.

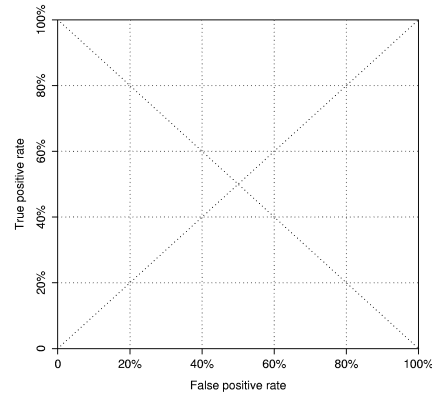


Figure 2. 2D ROC space.

Some of these special points and lines may have a slightly different interpretation in e.g. subgroup discovery or decision tree learning. For instance, the ascending diagonal here means a child node or subgroup that has the same class distribution as the parent node or overall population; the upper left (lower right) triangle contains subgroups deviating towards the positives (negatives). Therefore, metrics for subgroup discovery and splitting criteria are normally 0 on the ascending diagonal. The descending line $tpr+c*fpr=1$ represents subgroups whose size equals the number of positives in the population.

It is worth noting the following symmetries in contingency tables and ROC space. Exchanging columns in a

contingency table corresponds to, e.g., inverting the predictions of a classifier; in ROC space, this amounts to point-mirroring a point through (0.5,0.5). Exchanging rows in the contingency table, on the other hand, amounts to keeping the model the same but inverting the labels of the test instances; it corresponds to swapping true and false positive rates, i.e., line-mirroring ROC space across the ascending diagonal. Notice that this also affects the skew ratio (c becomes $1/c$). Exchanging both rows and columns, i.e., swapping the correct predictions for both classes as well as the misclassifications, corresponds to line-mirroring ROC space across the descending diagonal.

3. Machine Learning Metrics in ROC Space

This section is devoted to a geometric investigation of the behaviour of a range of metrics in ROC space. A more formal analysis is given in Section 4. As has been argued in the previous section, we consider metrics evaluating the performance of models in terms of their (estimated) true and false positive rates, which additionally take the skew ratio c as a parameter. Table 2 contains formulas for the main metrics considered in this paper. The formulas can be verified by substituting $c=NEG/POS$, $tpr=TP/POS$ and $fpr=FP/NEG$. Further explanation for these metrics is given in the corresponding subsection below.

Table 2. Metrics defined in terms of true and false positive rates and skew ratio. An asterisk indicates weak skew-insensitivity.

| METRIC | FORMULA | KEW-INSENSITIVE |
|------------|---|-----------------------------|
| ACCURACY | $\frac{tpr + c(1 - fpr)}{1 + c}$ | $\frac{(tpr + 1 - fpr)}{2}$ |
| PRECISION* | $\frac{tpr + c \cdot fpr}{2tpr}$ | $\frac{tpr + fpr}{2tpr}$ |
| F-MEASURE | $\frac{tpr + c \cdot fpr + 1}{4c}$ | $\frac{tpr + fpr + 1}{2}$ |
| WRACC* | $\frac{(1 + c)^2}{(1 + c)^2} (tpr - fpr)$ | $tpr - fpr$ |

3.1 Isometrics

We recall that *isometrics* are collections of points with the same value for the metric. Generally speaking, in 2D ROC space isometrics are lines or curves, while in 3D ROC space they are surfaces. If the isometric lines are independent of the skew ratio the isometric surfaces will be vertical; we will refer to such metrics as *strongly skew-insensitive*. Alternatively, the metric can be *weakly skew-insensitive* (isometric surfaces are non-vertical but their isometrics retain their shape when varying the skew ratio); or they can be *skew-sensitive* as in Figure 1. These concepts will be made more precise below.

We obtain isometric lines by fixing the skew ratio c . Most of the plots in this paper show isometrics for the unbiased case ($c=1$) as well as a biased case such as $c=1/2$ or $c=5$.

As explained previously, the parameter c tells us how positives and negatives should be traded off in the target context; we refer to it as the *expected skew ratio*, or briefly *expected skew*. We contrast this with the *effective skew ratio*, which is the slope of the isometric in a given point. This indicates the trade-off between true and false positive rates the metric makes locally in that point, which is important for determining the direction in which improvements are to be found. Below we will see three types of isometric landscapes: (a) with parallel linear isometrics (accuracy, WRACC); (b) with non-parallel linear isometrics (precision, F-measure); and (c) with non-linear isometrics (decision tree splitting criteria). Type (a) means that the metric applies the same positive/negative trade-off throughout ROC space; type (b) means that the trade-off varies with different values of the metric; and type (c) means that the trade-off varies even with the same value of the metric.

In addition, we will describe isometric landscapes associated with metric m using the following concepts: (1) The *tpr-indicator line*, where $m = tpr$. This is useful for reading off the value of the metric associated with a particular isometric. (2) The *line or area of skew-indifference*, which is the collection of points such that m is independent of c . This line may be a useful target if the expected skew is unknown, but known to be very different from the training set distribution. (3) The metric that results for $c=1$, which we call the *skew-insensitive version* of the metric (see the right column in Table 2).

3.2 Accuracy

It is obvious from Figure 1 that accuracy behaves differently in different slices of 3D ROC space, and thus is skew-sensitive (notice that the term ‘cost-sensitive’ might be considered confusing in this context). This can be explained by noting that accuracy ignores the distribution of correct predictions over the classes. Its definition is given in Table 2, and Figure 3 shows an isometric plot.

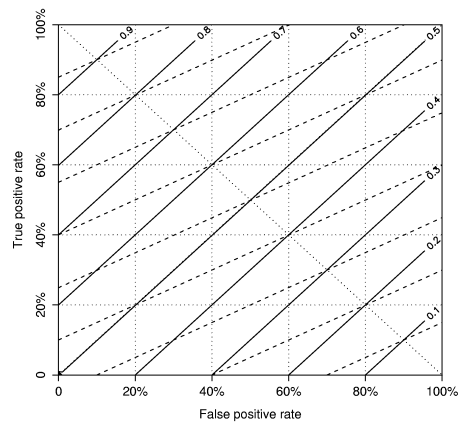


Figure 3. Accuracy isometrics for accuracy values 0.1, 0.2, ... 1. The solid lines with slope 1 result from an unbiased expected skew ratio ($c=1$), while the flatter dashed lines indicate an expected skew biased towards positives ($c=1/2$).

Not surprisingly, accuracy imposes a constant effective skew ratio throughout ROC space, which is equal to the expected skew ratio c . The flatter isometrics for $c=1/2$ indicate a bias towards performance on the positives. On the descending diagonal performance on positives and negatives is equal, and consequently unbiased and biased versions of accuracy give the same value — accuracy’s line of skew-indifference. In the lower left triangle (better performance on positives), biasing the expected skew ratio c towards negatives has the effect of decreasing accuracy in any fixed point, whereas in the upper right triangle the effect is opposite. The effect is larger the further we are from the descending diagonal. Finally, the tpr -indicator line for accuracy, where accuracy equals tpr , is the descending diagonal, independent of the expected skew. For $c=1$ accuracy reduces to $(tpr+1-fpr)/2$, which can be seen as a skew-insensitive version of accuracy (i.e., the average of true positive and true negative rates).

3.3 Precision

An entirely different isometric landscape is obtained when we plot precision, which is defined as $tpr/(tpr+c*fpr)$. Unlike accuracy, precision imposes a varying effective skew ratio: e.g., above the ascending diagonal effective skew is biased towards the negatives, and optimal precision is obtained if $fpr=0$. Precision has only trivial lines of skew-indifference (the fpr and tpr axes), but it is weakly skew-insensitive: we get the same isometrics for different values of c , only their associated values differ. A fully skew-insensitive version of precision is $tpr/(tpr+fpr)$. The descending diagonal is the tpr -indicator line for the skew-insensitive version of precision; in general the tpr -indicator line is given by $tpr+c*fpr=1$. Finally, notice that increasing precision’s bias towards positives by decreasing c results in increased precision values throughout ROC space.

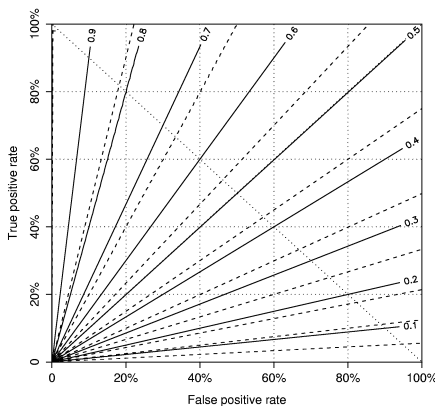


Figure 4. Precision isometrics for $c=1$ (solid lines) and $c=1/2$ (dashed lines).

3.4 F-measure

The *F-measure* (Van Rijsbergen, 1979) trades off precision= $TP/(TP+FP)$ and recall= $TP/(TP+FN)$ by averaging

FP and FN : $TP/(TP+Avg(FP,FN)) = 2TP/(2TP+FP+FN) = 2TP/(TP+FP+POS)$. This measure is insensitive to how the incorrect predictions are distributed over the classes. The F-measure can be rewritten as $2tpr/(tpr+c*fpr+1)$; a skew-insensitive version is $2tpr/(tpr+fpr+1)$. An isometric plot is given in Figure 5; because it is customary in information retrieval to have many more negatives than positives, the biased plot uses $c=5$.

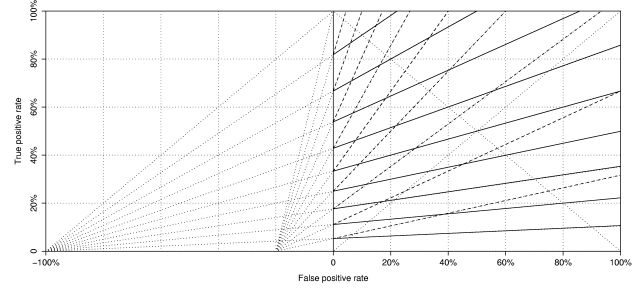


Figure 5. F-measure isometrics for $c=1$ (solid lines) and $c=5$ (dashed lines).

Figure 5 is drawn in this way to emphasise that the F-measure can be seen as a version of precision where the rotation point (for fixed c) is translated to the left to ($fpr=-1/c, tpr=0$). Obviously, when $c \gg 1$ the difference with precision becomes negligible (for that reason, a non-uniform weighting of false positives and false negatives is sometimes used). Also, notice how the tpr axis is a line of skew-indifference, where the F-measure has the same value regardless of the expected skew. Along this line the F-measure is a re-scaling of the true positive rate (i.e., $2tpr/tpr+1$); the tpr -indicator line is the same as for precision, i.e., $tpr+c*fpr=1$. Again, biasing c towards positives increases the value of the metric throughout ROC space.

It is interesting to note that versions of precision with shifted rotation point occur more often in machine learning. In (Fürnkranz & Flach, 2003) it is shown that both Laplace-corrected and m -corrected precision follow this pattern, and that by moving the rotation point further away from the origin we can approximate accuracy-like measures. Also, Gamberger and Lavrac (2002) use this device to control the generality of induced subgroups.

We propose the following simplification of the F-measure, which we call the *G-measure* = $TP/(FP+POS) = tpr/(c*fpr+1)$. This measure has the same isometrics as the F-measure, only its values are distributed differently. In particular, $fpr=0$ is both a line of skew-indifference and the tpr -indicator line for the G-measure. We prove the equivalence of F- and G-measures in Section 4.

3.5 Weighted Relative Accuracy

Subgroup discovery is concerned with finding subgroups of the population that are unusual with respect to the target distribution. *Weighted relative accuracy* (WRAcc) is a metric that compares the number of true positives with the expected number if class and subgroup were statistically

independent (Lavrac, Flach & Zupan, 1999): $tp - (tp + fp) \cdot pos = tp \cdot neg - fp \cdot pos$, which can be rewritten to $(c/(c+1))^2 \cdot (tpr - fpr)$ (if desired, it can be re-scaled to range from -1 to $+1$ by multiplication with 4). The isometrics are parallel to the ascending diagonal: WRAcc is essentially equivalent to an unbiased version of accuracy. However, it is only weakly skew-insensitive — a fully skew-insensitive version is $tpr - fpr$. We omit the isometric plot for lack of space.

It is worth pointing out the following connection with area under the ROC curve, which evaluates the aggregated quality of a set of classifiers (Hand & Till, 2001). This can be applied to a single model by combining it with the default models *AlwaysNegative* and *AlwaysPositive*: i.e., by constructing a three-point ROC curve $(0,0) - (fpr, tpr) - (1,1)$ from the origin, through a point (fpr, tpr) to the top-right corner. Geometrically, it is easy to see that the area under this curve is $(tpr - 1 + fpr)/2$, i.e., the average of true positive and true negative rates, which is the skew-insensitive version of accuracy; or $tpr - fpr$ if measured only above the ascending diagonal.

3.6 Decision Tree Splitting Criteria

Most decision tree splitting criteria compare the impurity of the unsplit parent with the weighted average impurity of the children. If we restrict attention to a binary split in a two-class problem, the split can be described by a contingency table where *POS* and *NEG* denote the positive and negative instances in the parent, true and false positives denote the instances in one child, and true and false negatives denote the instances in the other child. Impurity-based metrics then take the following form:

$$m = \text{Imp}(pos, neg) \square (tp + fp) \text{Imp}\left(\frac{tp}{tp + fp}, \frac{fp}{tp + fp}\right) \\ \square (fn + tn) \text{Imp}\left(\frac{fn}{fn + tn}, \frac{tn}{fn + tn}\right)$$

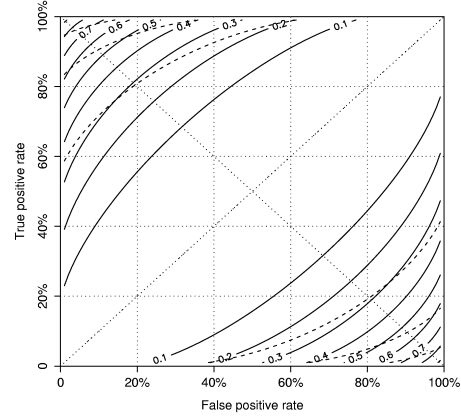
Impurity can be defined as entropy (Quinlan, 1986), Gini index (Breiman *et al.*, 1984), or DKM (Kearns & Mansour, 1996); their definitions can be found in Table 3. The right column of the table contains, for Gini index and DKM, the weighted impurity of the (tp, fp) child as a proportion of the impurity of the parent (*relative impurity*).

Table 3. Impurity measures used in decision tree splitting criteria (all scaled to $[0,1]$).

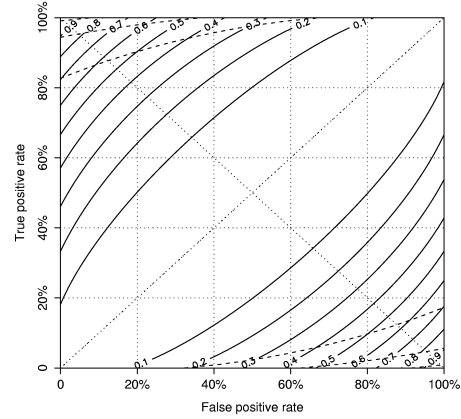
| IMPURITY | $\text{Imp}(p, n)$ | RELATIVE IMPURITY |
|----------|-------------------------------------|---|
| ENTROPY | $\square p \log p \square n \log n$ | |
| GINI | $4pn$ | $\frac{1+c}{tpr + c \cdot fpr} tpr \cdot fpr$ |
| DKM | $2\sqrt{pn}$ | $\sqrt{tpr \cdot fpr}$ |

Splitting criteria naturally have more symmetry than the metrics discussed in the previous sections. In particular, they are insensitive to swapping columns in the contin-

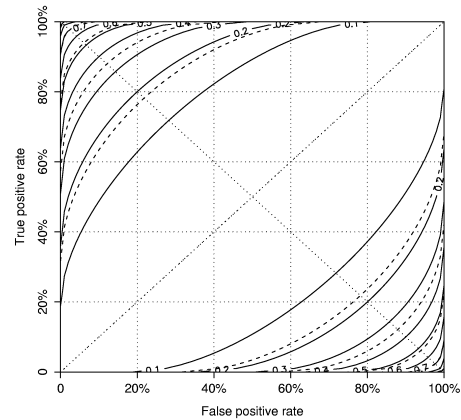
gency table: as has been pointed out before, this corresponds to point-mirroring through $(0.5, 0.5)$ in ROC space. The next three figures show isometrics for an impurity-based splitting criterion using these impurity measures. Because we want to investigate the influence of class skew, we plot isometrics for $c=1$ and $c=1/10$.



(a) Information gain



(b) Gini-split



(c) DKM-split

Figure 6. Splitting criteria isometrics for $c=1$ (solid lines) and $c=1/10$ (dashed lines).

The top two plots in Figure 6 demonstrate skew-sensitivity, as with decreasing c the isometrics become flatter; however, we can observe that Gini-split is much more dependent on the expected skew than information gain. DKM-split, on the other hand, is (weakly) skew-insensitive (Drummond & Holte, 2000): in ROC space, this amounts to having both the ascending and the descending diagonals as symmetry axes. The fact that DKM-split is skew-insensitive can also be seen from Table 3, as the relative impurity (the weighted impurity of the (tp, fp) child as a proportion of the impurity of the parent) is independent of c .

All these splitting criteria have non-linear isometrics: in points close to the edge of the plot, where one of the children is nearly pure, the effective skew is more biased. It is not entirely clear why this should be the case. For instance, in (Ferri, Flach & Hernandez, 2002) we report good results with an alternative splitting criterion that is not impurity-based but instead geared towards increasing the area under the ROC curve. The isometrics of this AUC-split metric can be easily shown to be straight parallel lines, but the metric is skew-sensitive.

The foregoing analysis gives rise to a novel, strongly skew-insensitive splitting criterion that can be derived from Gini-split by setting $c=1$. As can be observed from Table 3, this results in a relative impurity of the (tp, fp) child of $2tpr \cdot fpr / (tpr + fpr)$. Setting the impurity of the parent to 1, the complete formula is

$$\text{Gini-ROC} = 1 - \frac{2tpr \cdot fpr}{tpr + fpr} = \frac{2(1 - tpr)(1 - fpr)}{1 - tpr + 1 - fpr}$$

The idea is that the class distribution in the parent is taken into account when calculating the impurity of the children, through the true and false positive rates. There is no need for weighting the impurities of the children. The isometrics for this splitting criterion are equal to the Gini-split isometrics for $c=1$; they are flatter than the DKM-split isometrics, and thus place less emphasis on completely pure children. Generalisation of Gini-ROC to multiple classes and experimental comparison with DKM-split are interesting issues for further research. We would also like to mention that Gini-split can be shown to be equivalent to the χ^2 statistic normalised by sample size, up to a factor of $4c/(1+c)^2$, thus settling an open question in (Vilalta & Oblinger, 2000). It follows that Gini-ROC also establishes a strongly skew-insensitive version of χ^2 .

4. Formal Analysis

The main message of this paper is that machine learning metrics are distinguished by their effective skew landscapes rather than by their values per se. This is captured by the following definition.

Definition 1. Two metrics are *skew-equivalent* if they have the same effective skew ratio throughout 2D ROC space for each value of c .

In (Fürnkranz & Flach, 2003) we defined two metrics m_1 and m_2 to be equivalent if they are either compatible or antagonistic, where m_1 and m_2 are defined to be *compatible* if $m_1(x) > m_1(y)$ iff $m_2(x) > m_2(y)$ holds for all points x and y , and *antagonistic* if $m_1(x) > m_1(y)$ iff $m_2(x) < m_2(y)$. If two metrics are equivalent, they lead to the same results in any setting where the metric is used for ranking models. We now show that Definition 1 states a necessary condition for equivalence.

Theorem 1. *Equivalent metrics are skew-equivalent.*

Proof. Suppose m_1 and m_2 don't have the same effective skew ratio everywhere, then there exists a point x where their isometrics cross. Let y be another point close to x on the m_1 isometric, then $m_1(x) = m_1(y)$ but $m_2(x) \neq m_2(y)$, i.e., m_1 and m_2 are neither compatible nor antagonistic.

Skew-equivalence is not a sufficient condition for equivalence: tpr - fpr and $(tpr - fpr)^2$ are skew-equivalent but not equivalent. We conjecture that the difference between the two can be characterised by imposing continuity or monotonicity constraints on the metric.

We proceed to demonstrate equivalence of the F- and G-measures. Although it would be straightforward to prove compatibility, we prove skew-equivalence instead to demonstrate one technique for deriving an expression for the effective skew ratio of a metric.

Theorem 2. *The F- and G-measures are skew-equivalent.*

Proof. The isometric for a given value $tpr/(c \cdot fpr + 1) = m$ of the G-measure is given by $tpr = (mc \cdot fpr + m)$. The slope of this line is mc . In an arbitrary point (fpr, tpr) this is equal to $c \cdot tpr / (c \cdot fpr + 1)$. The isometric for a given value $2tpr / (tpr + c \cdot fpr + 1) = m$ of the F-measure is given by $tpr = (mc \cdot fpr + m) / (2 - m)$. The slope of this line is $mc / (2 - m)$. In an arbitrary point (fpr, tpr) this is again equal to $c \cdot tpr / (c \cdot fpr + 1)$.

We can use the notion of skew-equivalence to define (weak) skew-insensitivity.

Definition 2. A metric is *weakly skew-insensitive* if it is skew-equivalent to itself for all expected skew ratios. A metric is *strongly skew-insensitive* if it is identical to itself for all expected skew ratios.

Theorem 3. *Precision is weakly skew-insensitive.*

Proof. The isometric through any given point (fpr, tpr) goes through the origin and thus has slope tpr/fpr , which is independent of c .

The next result relates two metrics in terms of their effective skew ratios. This complements results from (Vilalta & Oblinger, 2000) who developed a method to quantify the difference between metrics.

Theorem 4. *For any expect skew ratio and throughout ROC space, the F-measure is more biased towards the positives than precision.*

Proof. The effective skew ratio of the F-measure in a

point (fpr, tpr) was derived in the proof of Theorem 2 as $c * tpr / (c * fpr + 1) < tpr / fpr$, which is the slope of the precision isometric through that point according to Theorem 3.

Table 4 gives a summary of the effective skew ratios of some of the metrics considered in this paper.

Table 4. Effective skew ratio (slope) of various metrics as a function of true and false positive rates and expected skew ratio.

| METRIC | EFFECTIVE SKEW RATIO |
|----------------------|----------------------|
| ACCURACY | c |
| PRECISION | tpr/fpr |
| F-MEASURE, G-MEASURE | $tpr/(fpr+1/c)$ |
| WRACC | 1 |

5. Concluding Remarks

In this paper we have proposed the use of ROC isometric plots to analyse machine learning metrics that are commonly used for model construction and evaluation. By deriving 2D ROC space from 3D ROC space we have made the central role of the skew ratio explicit. We have argued that the defining characteristic of a metric is its effective skew landscape, i.e., the slope of its isometric at any point in 2D ROC space. This provides a foundation on which a theory of machine learning metrics can be built. We have obtained a number of initial results, including a simplification of the F-measure commonly used in information retrieval, and compared metrics through their effective skew ratios. We have also demonstrated that, while both information gain and Gini-split are skew-sensitive, the latter is more so than the former. Finally, we have derived a skew-insensitive version of Gini-split as an alternative to the weakly skew-insensitive DKM-split.

There are a variety of ways in which this work can be taken forward, the most obvious of which is perhaps to include the purpose served by the metric in the analysis. For instance, an evaluation metric like accuracy considers each point in ROC space as a finished product, while search heuristics aim to find points that can be *turned into* finished products. We conjecture that the gradient of the metric, which is orthogonal to the slope of isometrics, would play an important role in this analysis. Another avenue for further research is to consider n by m contingency tables, for instance to obtain a multi-class and multi-split version of Gini-ROC.

Acknowledgements

I would like to express my gratitude to Johannes Fürnkranz for many enlightening discussions, as well as help with producing the isometric plots. Thanks are also due to the anonymous reviewers for many helpful comments. Support from National ICT Australia and the University of New South Wales, where I was a visiting research fellow during completion of this paper, is gratefully acknowledged.

References

- Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.
- Drummond, C., & Holte, R.C. (2000). Exploiting the cost (in)sensitivity of decision tree splitting criteria. *Proceedings of the 17th International Conference on Machine Learning (ICML-00)* (pp. 239–246). Morgan Kaufmann.
- Elkan, C. (2001). The foundations of cost-sensitive learning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)* (pp. 973–978). Morgan Kaufmann.
- Ferri, C., Flach, P., & Hernandez, J. (2002). Learning decision trees using the area under the ROC curve. *Proceedings of the 19th International Conference on Machine Learning (ICML-02)* (pp. 139–146). Morgan Kaufmann.
- Fürnkranz, J. & Flach, P. (2003). An analysis of rule evaluation metrics. *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*.
- Gamberger, D., & Lavrac, N. (2002). Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17, 501–527.
- Hand, D., & Till, R. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45, 171–186.
- Kearns, M., & Mansour, Y. (1996). On the boosting ability of top-down decision tree learning algorithms. *Proceedings ACM Symposium on the Theory of Computing* (pp. 459–468). ACM Press.
- Lavrac, N., Flach, P., & Zupan, B. (1999). Rule evaluation measures: A unifying view. *Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP-99)* (pp. 174–185). Springer-Verlag.
- Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. *Proceedings of EuroSpeech* (pp. 1895–1898).
- Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42, 203–231.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1, 81–106.
- Van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.
- Vilalta, R., & Oblinger, D. (2000). A quantification of distance-bias between evaluation metrics in classification. *Proceedings of the 17th International Conference on Machine Learning (ICML-00)* (pp. 1087–1094). Morgan Kaufmann.