Classification in Context: Adapting to changes in class and cost distribution

Peter Flach

Intelligent Systems Laboratory, University of Bristol, UK

Abstract. Two natural notions of context in predictive machine learning are given by knowledge about the distribution of the target variable on one hand, and costs of prediction errors on the other. In this paper I consider those two notions of context in binary classification. In the first part of the paper I explore the intuition that class and cost distributions are closely related, and show how changes in one can be handled in a similar way to changes in the other. Secondly, I investigate how a change of loss function affects these results, and propose an alternative to ROC space and associated cost space specifically designed for F-measure.

Keywords: ROC analysis, classification performance metrics, class distribution change, cost-sensitive classification, operating context.

1 Introduction and motivation

Two natural notions of context in predictive machine learning in general, and in binary classification in particular, are given by knowledge about the distribution of the target variable on one hand, and costs of prediction errors on the other. Such contexts often influence the chosen operating point of a classifier. For example, if false positives are very cheap then the always-positive classifier is hard to beat – this is the principle underlying spam e-mail. The same applies if negatives are rare. The influence of the cost context in binary classification was investigated by Hernández-Orallo et al. (2012). That paper considered different ways of setting decision thresholds for binary classifiers, where the more realistic of these would take a cost context into account when setting the threshold. By averaging over cost contexts it was possible to relate expected loss to well-known – non-cost-based – performance metrics.

Although class distributions and cost distributions can be seen as two sides of the same coin (Elkan, 2001), Hernández-Orallo et al. (2012) only focused on cost contexts. So one aim of this paper is to investigate how their results can be extended to cover changes in class distributions. As we will see, this is possible if we re-interpret their cost parameter c (the cost of misclassifying a positive in proportion to the cost of misclassifying both a positive and a negative) as a context *change* from training class distribution to deployment class distribution.

Secondly, we investigate what can also be seen as a context change: switching from an accuracy-based loss to an F-measure based loss, which behaves quite differently. The main change is then that we need to distinguish between true positives and true negatives because the latter are considered to have zero profit. We show that this still

can be captured by a univariate context parameter. We also argue that analysis is easier if we apply non-linear but monotonic transformations to true and false positive rate, and propose an alternative to ROC space and associated cost space specifically designed for F-measure. This can be seen as a step towards developing a similar theory linking threshold choice methods and performance metrics for loss functions based on F-measure.

The outline of the paper is as follows. In Section 2 we recall the main definitions and conceptual framework of Hernández-Orallo et al. (2012). Section 3 investigates how to extend these results to changes in class context. In Section 4 we initiate a similar analysis based on F-measure, and Section 5 concludes.

2 Background: cost contexts and performance metrics for binary classification

In this section we recall the main definitions and results of Hernández-Orallo et al. (2012).

2.1 Notation and definitions

A (single-label) *classifier* is a function that maps instances x from an instance space X to classes y from an output space Y. For this paper we will assume binary classifiers, i.e., $Y = \{0,1\}$. A *model* is a function $m : X \to \mathbb{R}$ that maps examples to real numbers (scores) on an unspecified scale. We use the convention that higher scores express a stronger belief that the instance is of class 1. A *probabilistic model* is a function $m : X \to [0,1]$ that maps examples to estimates $\hat{p}(1|x)$ of the probability of example x to be of class 1. In order to make predictions in the Y domain, a model can be converted to a classifier by fixing a decision threshold t on the scores. Given a predicted score $s = m(x)$, the instance x is classified in class 1 if $s > t$, and in class 0 otherwise.

For a given, unspecified model and population from which data are drawn, we denote the score density for class k by f_k and the cumulative distribution function by F_k. Thus, $F_0(t) = \int_{-\infty}^{t} f_0(s)ds = P(s \leq t|0)$ is the proportion of class 0 points correctly classified if the decision threshold is t, which is the sensitivity or true positive rate at t. Similarly, $F_1(t) = \int_{-\infty}^{t} f_1(s)ds = P(s \leq t|1)$ is the proportion of class 1 points incorrectly classified as 0 or the false positive rate at threshold t. Note that we use 0 for the positive class and 1 for the negative class, but scores increase with $\hat{p}(1|x)$. That is, $F_0(t)$ and $F_1(t)$ are monotonically non-decreasing with t. This has some notational advantages and is the same convention as used by, e.g., Hand (2009).

Given a dataset $D \subset \langle X, Y \rangle$ of size $n = |D|$, we denote by D_k the subset of examples in class $k \in \{0,1\}$, and set $n_k = |D_k|$. The (positive) *class proportion* or *class context* is then $\pi = n_0/n$.[1] Given a model and a threshold t, we denote by $R(t) = \pi F_0(t) + (1 - \pi)F_1(t)$ the predicted positive rate, i.e., the proportion of examples that will be predicted positive (class 0) if the threshold is set at t. Finally, the average score of actual class k is $\bar{s}_k = \int_0^1 s f_k(s)ds$.

[1] This is a slight deviation from Hernández-Orallo et al. (2012) who use $\pi_k = n_k/n$; in this paper subscripts indicate quantities that are not (necessarily) normalised over the classes.

2.2 Threshold choice methods and expected loss

A threshold choice method is a (possibly non-deterministic) function $T : \Theta \to \mathbb{R}$ such that given an operating context it returns a decision threshold. Given a threshold choice function T, the loss for a particular operating context θ is denoted by $Q(T(\theta);\theta)$. If we do not know the operating context in advance, we can define a distribution for operating contexts as a distribution $w(\theta)$, and calculate expected loss as a weighted average over operating contexts:

$$L \triangleq \int_{\Theta} Q(T(\theta);\theta)w(\theta)d\theta \tag{1}$$

In classification a typical loss function is the error rate, which can be defined in terms of class distribution and true and false positive rates as follows:

$$Q(t;\pi) \triangleq \pi(1-F_0(t)) + (1-\pi)F_1(t) \tag{2}$$

In cost-based classification we can denote the cost of misclassifying a positive as c_0 and of misclassifying a negative as c_1, leading to the following loss:

$$Q(t;\pi,c_0,c_1) \triangleq c_0\pi(1-F_0(t)) + c_1(1-\pi)F_1(t) \tag{3}$$

Setting $b \triangleq c_0 + c_1$ for the cost associated with misclassifying one positive and one negative, and $c \triangleq c_0/b$ for the relative cost of misclassifying a positive, we obtain the following alternative parametrisation:[2]

$$Q(t;\pi,b,c) \triangleq b\{c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)\} \tag{4}$$

Hernández-Orallo et al. (2012) argued that it makes sense in many situations to assume b and c independent. If we also assume π fixed this leads to the following expected loss:

$$L = \mathbb{E}\{b\} \int_0^1 \{c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)\}\, w(c)dc \tag{5}$$

This means that for the expected loss the variability in b is only captured through its expected value. We will further simplify the loss function by fixing $b = 2$ which means that loss at any threshold is commensurate with error rate.

So in this setting the context is given by the relative cost: $c = 1/2$ means that we give misclassified positives and negatives the same weight, as in error rate; $c > 1/2$ means that misclassified positives (false negatives) cost more than misclassified negatives (false positives); and $c < 1/2$ means that false positives are more expensive than false negatives. One way to interpret c is as a cost-based deployment context, relative to uniform misclassification costs in training. This is a common scenario in machine learning, where we use error rate as the loss function in training, and adapt the model

[2] Negative costs for correct classifications can easily be taken into account: in that case b is the total cost associated with one true positive, one false positive, one false negative and one true negative; and c is the relative cost of one true positive and one false negative.

to non-uniform misclassification costs in deployment by setting an appropriate threshold. Expected loss then quantifies the performance of this deployment strategy over a distribution of cost contexts. In particular, Hernández-Orallo et al. (2012) investigated expected loss over uniform cost contexts for a range of threshold choice methods (their main technical results are repeated in an Appendix for convenience).

3 Re-interpreting c as a class distribution context change

The results of Hernández-Orallo et al. (2012) were specifically derived for varying cost contexts, parametrised by the relative cost c of misclassifying a positive. At this point a reader might say: 'I do not deal with cost-sensitive classification, for me it always holds that $c = 1/2$ and so only the results that fix the threshold or ignore the cost context are of any relevance to me'. This is a valid point – yet it seems equally intuitive that we may want to adapt the decision threshold to changes in class distribution. So the aim of this section is to investigate to what extent the previous results can be extended to such changes in class context.

3.1 The effect of changing the prior

One of the first papers considering this question is by Elkan (2001). He considers a change of positive prior from π to z, and answers the question: if p is the correct posterior probability for an instance for prior π, and assuming the class-conditional likelihoods remain unchanged, what is the adjusted posterior probability p' for the new prior z? The following formula gives the answer:

$$p' = z\frac{p - p\pi}{p - p\pi + zp - \pi z} \tag{6}$$

In particular, consider an instance on the decision boundary, i.e. $p = 1/2$, and let's write c for the decision threshold corresponding to the new class prior z. We have

$$c = z\frac{(1-\pi)/2}{(1-\pi)/2 + z(1/2-\pi)} = \frac{(1-\pi)z}{(1-\pi) + z(1-2\pi)} = \frac{(1-\pi)z}{(1-\pi)z + \pi(1-z)} \tag{7}$$

The relation between c and z is invertible:

$$z = \frac{\pi c}{\pi c + (1-\pi)(1-c)} \tag{8}$$

This quantity was actually considered by Hernández-Orallo et al. (2012) as an operating context combining both class and cost distributions; they called it *skew*. The corresponding loss is

$$Q(t;z) \triangleq \{z(1 - F_0(t)) + (1-z)F_1(t)\} \tag{9}$$

Note that if $\pi = 1/2$ then $z = c$: this allows to easily generalise the results for uniform c to uniform z.

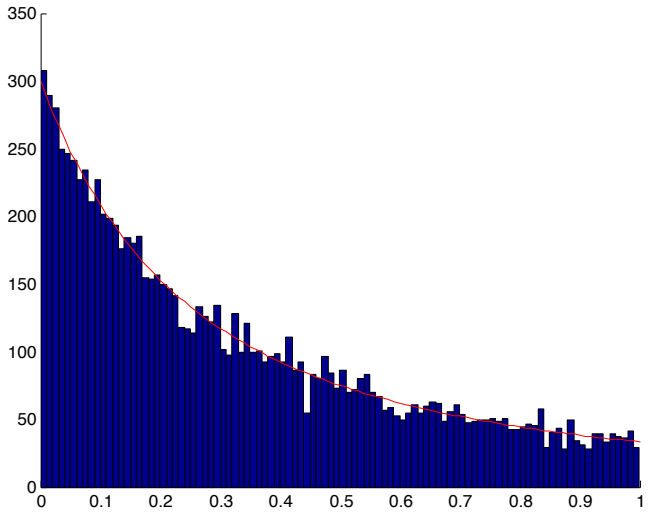

Fig. 1. The blue histogram shows the distribution of z-values that were obtained using $z = \phi_\pi(c)$ and $\pi = 0.25$ with 10 000 uniformly sampled c-values. The red line shows the pdf which according to Lemma 1 is $\frac{\pi(1-\pi)}{((1-\pi)z+\pi(1-z))^2}$ (rescaled to match the histogram).

The above analysis shows that we if we reinterpret z as deployment class prior, then c denotes the *change* in decision threshold from $1/2$. Denote the mapping from c to z as ϕ_π, then $\psi_\pi(z) = \phi_\pi^{-1}(z)$ calculates context change c from training prior π and deployment prior z. For example, if at training time we have balanced classes ($\pi = 1/2$) while at deployment time we have 20% positives ($z = 0.2$) then the *context change* $c = 0.2$ expresses that at deployment negatives are much more important compared to training. One way in which we could respond to this when considering probabilistic classifiers is to lower the decision threshold from 0.5 to 0.2. This has the desired effect of making more negative predictions (recall that scores are expected to be low for the positive class). Conversely, if $\pi = 0.2$ while at deployment classes are balanced ($z = 1/2$) then the score-driven threshold would be $c = 0.8$ to reflect the fact that we need to give positives more weight in deployment as compared to training. The same context change links, for example, $\pi = 0.4$ and $z = 0.73$.

3.2 Expected loss under uniform class context changes

The upshot of this is that we can interpret the results of Hernández-Orallo et al. (2012) without reference to cost-sensitive classification, by treating c not as a cost context but as a change in class distribution from training to deployment. The default option is that there is no context change, expressed by $c = 1/2$: in that case the score-driven threshold choice method prescribes that we should use the standard threshold of $1/2$ on the posterior probability, and the rate-driven threshold choice method prescribes that we should split our ranking in half. These default options can be adapted to actual context changes by deriving c from the training class context π and the deployment class context

z. The expected loss is then calculated over uniform context changes. Since there is a bijection between c and z, any probability distribution over c can be translated into a probability distribution over z (and vice versa), as expressed by the following Lemma.

Lemma 1. *If $p_c(c)$ is a probability density function over c, then*

$$p_z(z) = p_c(\psi_\pi(z)) \frac{\pi(1-\pi)}{((1-\pi)z+\pi(1-z))^2}$$

is the corresponding density over z.

In particular, if c is uniformly distributed then the pdf of z is equal to the right-hand term above (which is the derivative of $\psi_\pi(z)$) and the cumulative distribution function is $\psi_\pi(z)$ with median $z = \pi$. Figure 1 shows an example.

Formally, if $w(z)$ is a distribution over deployment class priors then expected (cost-insensitive) loss is given as:

$$L_z = \int_0^1 Q_z(T(z))w(z)dz \tag{10}$$

$$Q_z(t) = z(1-F_0(t)) + (1-z)F_1(t) \tag{11}$$

If we take $w(z)$ to be $\psi'_\pi(z)$ then we can apply the change of variable $z = \phi_\pi(c)$:

$$L_z = \int_0^1 Q_z(T(\phi_\pi(c)))dc \tag{12}$$

$T(\phi_\pi(c))$ depends on the threshold choice method: for example, for the score-driven threshold choice method we have $T(z) = \psi_{pi}(z)$ and hence $T(\phi_\pi(c)) = c$. The important point is that we are now uniformly integrating over context changes c, which means that we can re-interpret the results of Hernández-Orallo et al. (2012) as follows.

Fixed score thresholds. If a classifier sets the decision threshold at a fixed value irrespective of the operating context or the model – as happens, for example, with any classifier that applies a majority-class decision rule or a probabilistic classifier that sets a fixed threshold of $1/2$ on the estimated posterior probability – and $\mathbb{E}\{c\} = 1/2$ (i.e., on average the class distribution is not expected to change from training to deployment) then expected loss is equal to the error rate at that decision threshold. So the use of error rate as performance metric can be justified if we assume fixed decision thresholds and no context change on average. Notice that this result only depends on the expected value of c and not on the distribution. Furthermore, if $\mathbb{E}\{c\}$ differs from $1/2$ then we replace error rate with an appropriately weighted average of false positive rate and false negative rate.

Score-uniform thresholds and stochastic classifiers. Assuming probabilistic scores, a uniformly randomly chosen decision threshold, and no context change on average, expected loss is equal to the model's mean absolute error $MAE = \pi\bar{s}_0 + (1-\pi)(1-\bar{s}_1)$. One may ask why setting a random decision threshold makes sense, but notice that this

is equivalent, in expectation, to not setting a threshold at all but using the score to make stochastic predictions: positive with probability $1-s$ and negative with probability s. So the use of mean absolute error as performance metric can be justified if we assume a stochastic classifier and no context change on average. Again, if $\mathbb{E}\{c\}$ differs from $1/2$ then we re-weight the two terms in *MAE*.

Score-driven thresholds. The previous two cases ignored the context change c when setting the decision threshold and hence only depended on its expected value when averaging over context changes to obtain expected loss. The *score-driven* threshold choice method sets the threshold of a probabilistic classifier equal to c; in order to calculate expected loss we need to assume a particular distribution over c. Hernández-Orallo et al. (2011) obtained the following result: assuming probabilistic scores and a decision threshold equal to the context change c, expected loss under a uniform distribution of context changes is equal to the model's Brier score $BS = \pi \int_0^1 s^2 f_0(s)ds + (1-\pi)\int_0^1 (1-s)^2 f_1(s)ds$. Clearly the Brier score is never larger and usually smaller than mean absolute error, so this demonstrates the benefits of knowing and utilising the deployment context.

Rate-uniform thresholds and rank-stochastic classifiers. The next two results use ranks rather than scores to make predictions. Given a set of n instances we can achieve any rate $r/n, 1 \leq k \leq n$ by setting the threshold between the instances with rank r and $r+1$. Assuming thresholds are set to achieve a uniform random rate, expected loss decreases linearly with the *AUC* of the model. This result was first proved for uniform cost contexts by Flach et al. (2011), providing the first published justification for the use of ranking performance as a classification metric. It was later generalised to non-uniform cost contexts – in fact, the expected value of c only appears as an additive constant in the expected loss. Again, one may ask why setting a random rate makes sense, but notice that this is equivalent, in expectation, to using the ranks $r \in \{1,\ldots,n\}$ to make stochastic predictions: positive with probability $\frac{n-r}{n-1}$ and negative with probability $\frac{r-1}{n-1}$.

Rate-driven thresholds. Assuming thresholds are set to achieve a predicted positive rate equal to the context change c, expected loss for uniform context changes is again linearly related to *AUC* and in fact $1/6$ below the rate-uniform case, regardless of the model. This is a more realistic justification for using *AUC* as a classification performance metric. The constant loss reduction can be understood by noting that even for a random (or worse than random) model we can achieve near-zero loss for extreme context changes by always predicting the most heavily weighted class.

Optimal thresholds. Finally, we can set thresholds to directly minimise the loss Q. If we set thresholds optimally, the expected loss under uniform context changes is equal to the refinement loss corresponding to the model's convex hull. The refinement loss is the irreducible part of the model's Brier score (the remainder being the calibration loss). It is also equal to the area under the lower envelope of the model's cost lines (Drummond and Holte, 2006).

In conclusion, in this section I have argued that the results of Hernández-Orallo et al. (2012) bear relevance beyond the cost-sensitive setting in which they were formulated. The key notion here is the concept of a change in class distribution c, which can be obtained from training and deployment class proportions as in Equation 7. This leads to a more realistic non-uniform distribution over deployment class proportions which has the training proportion as its median. If a model is a good ranker but a poor probability estimator then we could set the threshold such that the predicted positive rate is equal to the operating context, with the ranking error as a proxy for expected loss under uniform context changes. If a model has decent probability estimates then we would set the threshold directly equal to c and use the Brier score as an estimator for expected loss under uniform context changes. A third possibility involves estimating the model's ROC convex hull and deriving thresholds that are optimal for specific ranges of the operating context. The corresponding performance metric is the model's refinement loss after 'convexification'. If the model is perfectly calibrated, the last two approaches yield the same results.

4 Alternative loss functions: the F-measure

The results so far are derived using an accuracy-based loss function, repeated here for convenience:

$$Q(t;\pi,b,c) = b\{c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)\} \tag{13}$$

Setting this equal to some constant q and solving for F_0 gives the equation for an accuracy-based loss isometric:

$$F_0(q;\pi,b,c) = \frac{1-\pi}{\pi}\frac{1-c}{c}F_1(q;\pi,b,c) + 1 - \frac{q/b}{c\pi} \tag{14}$$

It follows that these isometrics have constant slope $\frac{1-\pi}{\pi}\frac{1-c}{c}$, implying that the trade-off between true and false positive rates is constant everywhere and only dependent on the operating context. In this section we investigate what happens if we switch to a different loss function with different isometrics.

4.1 The F-measure and its isometrics

The F-measure has been introduced as an alternative for accuracy in order to deal with situations with many more negatives than positives but true negatives do not add value. For example, in information retrieval true negatives are non-answers to a query that are correctly not returned. If these carry value then it will be very hard to beat the always-negative classifier that will never return an answer to any query. Two measures that ignore the true negatives are precision $prec = TP/(TP+FP)$ and recall $rec = TP/(TP+FN)$; a combined measure can be obtained by averaging the false positives and false negatives, leading to the *F-measure* as the harmonic mean of precision and recall:

$$FM \triangleq \frac{TP}{TP+(FP+FN)/2} = \frac{2TP}{(TP+FP)+(TP+FN)} = \frac{2}{1/prec+1/rec} = \frac{2prec\cdot rec}{prec+rec} \tag{15}$$

Recall increases with increasing decision threshold (recall that lower scores indicate a stronger belief that the positive class applies), while precision can vary non-monotonically but converges to the positive class proportion π at total recall. This is often visualised in precision-recall plots with recall on the x-axis and precision on the y-axis.

In terms of true and false positive rate F-measure is defined as follows:

$$FM(t) = \frac{2\pi F_0(t)}{2\pi F_0(t) + \pi(1-F_0(t)) + (1-\pi)F_1(t)} = \frac{2\pi F_0(t)}{\pi(1+F_0(t)) + (1-\pi)F_1(t)} \quad (16)$$

The corresponding loss is then

$$1 - FM(t) = \frac{\pi(1-F_0(t)) + (1-\pi)F_1(t)}{2\pi F_0(t) + \pi(1-F_0(t)) + (1-\pi)F_1(t)} = \frac{\pi(1-F_0(t)) + (1-\pi)F_1(t)}{\pi(1+F_0(t)) + (1-\pi)F_1(t)} \quad (17)$$

F-measure isometrics in ROC space are straight lines with varying slope, rotating around the (virtual) point $(F_1 = -\pi/(1-\pi), F_0 = 0)$. We proceed to derive a scalar operating condition that can be interpreted as cost context or class context change, as before.

Assume the weight of a true positive is 1, and the weights of false positives and false negative sum up to $b = 2$ as before, then the F-measure loss can be expressed as

$$\begin{aligned} FQ(t;c) &\triangleq \frac{c_0\pi(1-F_0(t)) + c_1(1-\pi)F_1(t)}{2\pi F_0(t) + c_0\pi(1-F_0(t)) + c_1(1-\pi)F_1(t)} \\ &= \frac{c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)}{\pi F_0(t) + c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)} \end{aligned} \quad (18)$$

Setting this equal to some value q and solving for F_0 gives the equation of an F-measure isometric:

$$F_0(q;c) = \frac{1-\pi}{\pi}\frac{(1-c)(1-q)}{c+(1-c)q}F_1(q;c) + \frac{c(1-q)}{c+(1-c)q} \quad (19)$$

We see that the slope of an F-measure isometric depends on the loss $FQ = q$ and hence is not constant throughout ROC space: the added value of an increased true positive rate or a decreased false positive rate depends on the model's location. This is visualised in Figure 2 (left) for $c = 1/2$, leading to isometrics with slope $\frac{1-\pi}{\pi}\frac{1-q}{1+q}$. We see that increasing true positive rate is very highly valued at low recall, whereas decreasing false positive rate is very highly valued close to ROC heaven.

Example 1.1 (Determining dominance). Consider two classifiers A and B, such that B's true and false positive rate are 10% and 20% higher than A's, respectively. Without any further information about the actual true and false positive rates we can conclude that they have identical accuracy-based loss in any operating context where $\frac{1-\pi}{\pi}\frac{1-c}{c} = 1/2$.

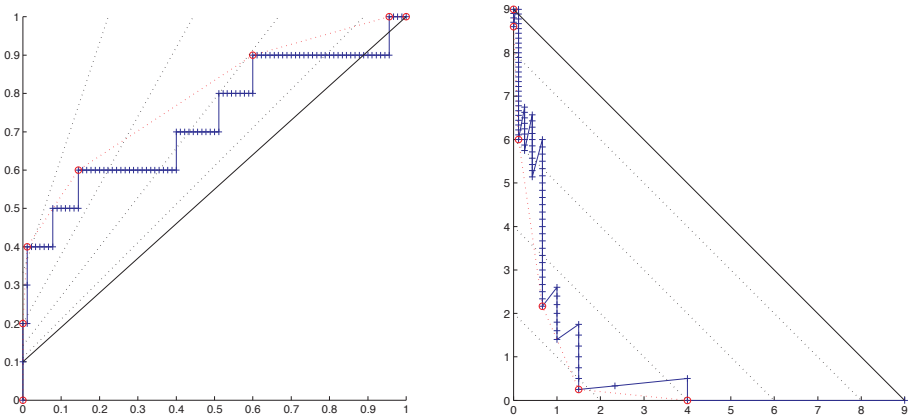

Fig. 2. (left) Example empirical ROC curve with F-measure isometrics. The ROC-convex hull consists of the seven points indicated in red. There are 10 positives and 90 negatives so the isometrics rotate around $F_1 = -1/9$. From top to bottom the F-measure values follow the harmonic series 1 (through ROC heaven), $1/2$, $1/3$, $1/4$ and $1/5$. The solid isometric corresponds to the default all-positive classifier **(right)** Corresponding FROC plot with $G_y = 1/prec - 1 = FP/TP$ on the y-axis and $G_x = 1/rec - 1 = FN/TP$ on the x-axis. FROC heaven is in the origin and F-measure isometrics are parallel lines with slope -1. The plot is bounded by the all-positive isometric so both axes can finish at $(1-\pi)/\pi$. The FROC-convex hull is formed by six of the ROCCH points (the seventh point is a singularity for F-measure).

If we are now told that in fact $c = 1-\pi$ then we can conclude that A dominates B with regard to accuracy-based loss in this context. However, for F-measure loss we need more information in order to determine dominance. For example, if $c = 1/2$ then Equation 19 tells us that F-measure isometrics have slope $\frac{1-q}{1+q}$, from which we can only conclude that A is the better classifier if its F-measure loss is less than $1/3$.

It is not hard to see that an F-measure isometric is never steeper than an accuracy isometric, with equality obtained only in ROC heaven ($q = 0$ in Equation 19). This implies that the optimal operating point for F-measure is always to the left of the optimal operating point for accuracy (on a non-ideal ROC curve), in accordance with results by Zhao et al. (2013); Chase Lipton et al. (2014). This implies that, while score-driven thresholds are optimal for accuracy if the model is calibrated, they are not optimal for F-measure. This invites a systematic investigation of possible threshold choice methods for F-measure loss and associated performance measures, building on the work by Dembczynski et al. (2011). While such an investigation is left for future work in this paper, I will suggest in what follows that a change of coordinates will be beneficial, leading to the notion of FROC curves and F-cost curves as direct analogues for the accuracy-based ROC curves and cost curves.

4.2 FROC curves

Equation 18 shows that FQ depends non-linearly on F_0 and F_1, but Equation 19 shows that for every fixed value of FQ there is a linear relationship between F_0 and F_1. I will now derive non-linear transformations of FQ, F_0 and F_1 that are linearly related and hence lead to constant-slope isometrics. Rewriting Equation 15 suggests how:

$$\frac{2}{FM} = \frac{1}{prec} + \frac{1}{rec} \tag{20}$$

Since $1/prec$ and $1/rec$ run from 1 to ∞ we subtract 1, which means that on the x-axis we have $G_x = 1/rec - 1 = FN/TP = (1-F_0)/F_0$ and on the y-axis we have $G_y = 1/prec - 1 = FP/TP = (1-\pi)F_1/\pi F_0$. The result is shown in Figure 2 (right). We then have that

$$G_x + G_y = \frac{1}{rec} - 1 + \frac{1}{prec} - 1 = \frac{2}{FM} - 2 = 2\frac{1-FM}{FM} = 2\frac{FQ}{1-FQ} \tag{21}$$

Note that $2\frac{FQ}{1-FQ}$ is a monotonic transformation of FQ, similar to a transformation of probabilities into odds. The factor 2 arises because F-measure is the harmonic mean (rather than the harmonic sum) of precision and recall.

F-measure isometrics have slope -1 in this plot, but this is without considering contexts. In order to incorporate contexts we use Equation 18 to get

$$2\frac{FQ}{1-FQ} = 2\frac{c\pi(1-F_0) + (1-c)(1-\pi)F_1}{\pi F_0} = 2cG_x + 2(1-c)G_y \tag{22}$$

from which we see that F-context isometrics have slope $-\frac{c}{1-c}$.

So an *FROC curve* is a plot of $G_y(s) = (1-\pi)F_1(s)/\pi F_0(s)$ on the y-axis against $G_x(s) = (1-F_0(s))/F_0(s)$ on the x-axis. The plot is bounded by the all-positive classifier which has $FQ = \frac{(1-c)(1-\pi)}{\pi+(1-c)(1-\pi)}$ and hence $2FQ/(1-FQ) = (1-\pi)/\pi$, which bounds both axes. We can calculate the slope of the FROC curve as follows:

$$g_x(s) = \frac{dG_x(s)}{ds} = -\frac{f_0(s)}{F_0(s)^2} \tag{23}$$

$$g_y(s) = \frac{dG_y(s)}{ds} = \frac{(1-\pi)}{\pi}\frac{f_1(s)F_0(s) - F_1(s)f_0(s)}{F_0(s)^2} \tag{24}$$

$$\frac{g_y(s)}{g_x(s)} = -\frac{(1-\pi)}{\pi}\left[\frac{f_1(s)}{f_0(s)}F_0(s) - F_1(s)\right] \tag{25}$$

At an optimal threshold t_F this slope is equal to $-\frac{c}{1-c}$, which gives

$$\frac{f_1(t_F)}{f_0(t_F)}F_0(t_F) - F_1(t_F) = \frac{c\pi}{(1-c)(1-\pi)} \tag{26}$$

Notice that $f_0(s)/f_1(s)$ is the slope of the ROC curve. For a calibrated classier $\pi f_0(s)/(1-\pi)f_1(s) = (1-s)/s$, from which we can derive

$$t_F = \frac{c\pi + (1-c)(1-\pi)F_1(t_F)}{c\pi + (1-c)(\pi F_0(t_F) + (1-\pi)F_1(t_F))} \tag{27}$$

In contrast, the optimal threshold for accuracy-based loss is $t_A = c$, assuming the classifier is calibrated.

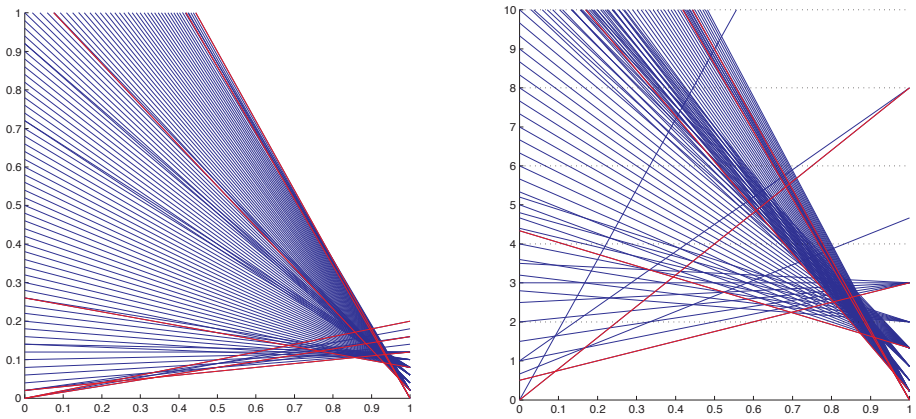

Fig. 3. (left) Accuracy-based cost lines. The x-axis shows c and the y-axis shows accuracy-based loss. The red cost lines correspond to operating points on the ROC convex hull. **(right)** Cost lines for F-measure loss. The y-axis shows $2FQ/(1-FQ)$. We can see that the optimal operating points are chosen for lower values of c, as expected.

4.3 F-cost curves

In an accuracy-based cost plot we transform a ROC operating point (F_1, F_0) into a cost line running from $2(1-\pi)F_1$ for $c=0$ on the left to $2\pi(1-F_0)$ for $c=1$ on the right (Figure 3 (left)). An ascending cost line ($(1-\pi)F_1 < \pi(1-F_0)$) indicates an operating point which has fewer false positives than false negatives, and so increasing c puts more weight on the incorrectly classified positives (i.e., false negatives) which increases the loss. Conversely, a descending cost line indicates more false positives than false negatives, and hence increasing c decreases the loss. The area under the lower envelope of the cost lines is the expected loss for uniform c if thresholds are chosen optimally, which corresponds to the refinement loss of the ROC convex hull.

We can similarly construct a cost plot for the F-measure which runs from $2G_y = 2(1-\pi)F_1/\pi F_0$ on the left to $2G_x = 2(1-F_0)/F_0$ on the right (Figure 3 (right)). We still have that these lines are ascending if $(1-\pi)F_1 < \pi(1-F_0)$ and descending if $(1-\pi)F_1 > \pi(1-F_0)$ but the cost lines cross at different values of c. The lower envelope now quantifies $\mathbb{E}[2FQ/(1-FQ)]$ for uniform c if thresholds are chosen optimally for F-measure loss.

5 Conclusions and further work

Two natural notions of context in predictive machine learning are given by knowledge about the distribution of the target variable on one hand, and costs of prediction errors on the other. In this paper we have considered those two notions of context in binary classification. In the first part of the paper we have explored the intuition that class and cost distributions are closely related, and that changes in one can be handled in a similar way to changes in the other. Specifically, I have shown that the cost context

c in Hernández-Orallo et al. (2012) can be re-interpreted as a class context change, thereby widening the relevance of the results of that paper. Secondly, we have started an investigation into how a change of loss function affects these results. Specifically, I have argued that for F-measure a transformation into an alternative to ROC space facilitates the analysis.

There are many avenues for further work, among which I mention two on which we are currently working. One direction concerns threshold choice methods specifically designed for optimising F-measure, and how they relate to models that output well-calibrated posterior probabilities. Secondly, there are many settings including multi-label and muti-class classification in which contexts are naturally multivariate, and it will be useful to extend the study of threshold choice methods and evaluation metrics to those multivariate contexts.

Acknowledgements

Comments from the anonymous reviewers helped improve this paper and are gratefully acknowledged. This work was partly supported by the REFRAME project granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA), and funded by the Engineering and Physical Sciences Research Council in the UK under grant EP/K018728/1.

Appendix: Summary of previous results for cost contexts

For reference I give the main results of Hernández-Orallo et al. (2012) and related papers (Hernández-Orallo et al., 2011; Flach et al., 2011).

Proposition 1 (Hernández-Orallo et al., 2012). *If a classifier sets the decision threshold at a fixed value t irrespective of the operating context or the model, then expected loss for any cost context distribution w is given by:*

$$L^{sf}(t) = 2\left\{\mathbb{E}_w\{c\}\pi(1-F_0(t)) + (1-\mathbb{E}_w\{c\})(1-\pi)F_1(t)\right\}.$$

Corollary 1. *If a classifier sets the decision threshold at a fixed value irrespective of the operating context or the model and $\mathbb{E}_w\{c\} = 1/2$ – i.e., on average there is no context change from training to deployment – then expected loss is equal to the error rate at that decision threshold.*

$$L^{sf}_{\mathbb{E}\{c\}=1/2}(t) = \pi(1-F_0(t)) + (1-\pi)F_1(t) = 1 - Acc(t).$$

Proposition 2 (Hernández-Orallo et al., 2012). *Assuming probabilistic scores and a uniformly randomly chosen decision threshold, expected loss under a distribution of cost contexts w is equal to:*

$$L^{su} = 2\left\{\mathbb{E}_w\{c\}\pi\bar{s}_0 + (1-\mathbb{E}_w\{c\})(1-\pi)(1-\bar{s}_1)\right\}.$$

Corollary 2. *Assuming probabilistic scores, uniform random thresholds or stochastic predictions, and no context change on average, expected loss is equal to the model's mean absolute error.*

$$L^{su}_{\mathbb{E}\{c\}=1/2} = \pi\bar{s}_0 + (1-\pi)(1-\bar{s}_1) = MAE.$$

Proposition 3 (Hernández-Orallo et al., 2011). *Assuming probabilistic scores and a decision threshold equal to the cost context c, expected loss under a uniform distribution of cost contexts is equal to the model's Brier score*

$$L^{sd}_{U(c)} = \pi\int_0^1 s^2 f_0(s)ds + (1-\pi)\int_0^1 (1-s)^2 f_1(s)ds = BS.$$

Proposition 4 (Hernández-Orallo et al., 2012). *Assuming thresholds are set to achieve a uniform random rate, expected loss under a distribution of cost contexts w decreases linearly with AUC as follows:*

$$L^{ru} = \pi(1-\pi)(1-2AUC) + \pi\mathbb{E}_w\{c\} + (1-\pi)(1-\mathbb{E}_w\{c\}).$$

Corollary 3. *Assuming thresholds are set to achieve a uniform random rate (or rank-stochastic predictions), and no context change on average, expected loss decreases linearly with AUC as follows:*

$$L^{ru}_{\mathbb{E}\{c\}=1/2} = \pi(1-\pi)(1-2AUC) + 1/2.$$

Proposition 5 (Hernández-Orallo et al., 2012). *Assuming thresholds are set to achieve a predicted positive rate equal to the cost context c, expected loss for uniform cost contexts is linearly related to AUC as follows:*

$$L^{rd}_{U(c)} = (1-\pi)\pi(1-2AUC) + 1/3 = L^{ru}_{\mathbb{E}\{c\}=1/2} - 1/6$$

Definition 1. *The optimal threshold choice method is defined as:*

$$T^o(c) \triangleq \arg\min_t \{Q(t;c)\} = \arg\min_t 2\{c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)\} \quad (28)$$

Proposition 6 (Hernández-Orallo et al., 2012). *If we set thresholds optimally, the expected loss under uniform cost contexts is equal to the refinement loss corresponding to the model's convex hull.*

$$L^o_{U(c)}(m) = RL(\mathrm{Conv}(m)).$$

Proposition 7 (Hernández-Orallo et al., 2012). *$BS = CL + RL$, where calibration loss CL and refinement loss CL are defined as follows:*

$$CL = \int_0^1 \frac{(s(\pi f_0(s) + (1-\pi)f_1(s)) - (1-\pi)f_1(s))^2}{\pi f_0(s) + (1-\pi)f_1(s)} ds.$$

$$RL = \int_0^1 \frac{(1-\pi)f_1(s)\pi f_0(s)}{\pi f_0(s) + (1-\pi)f_1(s)} ds.$$

Bibliography

Z. Chase Lipton, C. Elkan, and B. Narayanaswamy. Thresholding Classifiers to Maximize F1 Score. *ArXiv e-prints*, February 2014.

Krzysztof J Dembczynski, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. An exact algorithm for F-measure maximization. In *Advances in neural information processing systems*, pages 1404–1412, 2011.

C. Drummond and R. C. Holte. Cost curves: an improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

C. Elkan. The foundations of cost-sensitive searning. In *Proceedings of the seventeenth International Conference on Artificial Intelligence (IJCAI-01)*, pages 973–978, San Francisco, CA, 2001.

P. A. Flach, J. Hernández-Orallo, and C. Ferri. A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning, ICML2011*, 2011.

D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009.

J. Hernández-Orallo, P. A. Flach, and C. Ferri. Brier curves: a new cost-based visualisation of classifier performance. In *Proceedings of the 28th International Conference on Machine Learning, ICML2011*, 2011.

José Hernández-Orallo, Peter Flach, and Cesar Ferri. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869, 2012.

Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond Fano's inequality: bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *The Journal of Machine Learning Research*, 14(1):1033–1090, 2013.