Brier Curves: A New Cost-Based Visualisation of Classifier Performance

José Hernández-Orallo JORALLO@DSIC.UPV.ES
Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain

Peter Flach PETER.FLACH@BRISTOL.AC.UK
Intelligent Systems Laboratory, University of Bristol, UK

Cèsar Ferri CFERRI@DSIC.UPV.ES
Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain

Abstract

It is often necessary to evaluate classifier performance over a range of operating conditions, rather than as a point estimate. This is typically assessed through the construction of 'curves' over a 'space', visualising how one or two performance metrics vary with the operating condition. For binary classifiers in particular, cost space is a natural way of showing this range of performance, visualising loss against operating condition. However, the curves which have been traditionally drawn in cost space, known as *cost curves*, show the *optimal* loss, and hence assume knowledge of the optimal decision threshold for a given operating condition. Clearly, this leads to an optimistic assessment of classifier performance. In this paper we propose a more natural way of visualising classifier performance in cost space, which is to plot probabilistic loss on the *y*-axis, i.e., the loss arising from the probability estimates. This new curve provides new ways of understanding classifier performance and new tools to compare classifiers. In addition, we show that the area under this curve is exactly the Brier score, one of the most popular performance metrics for probabilistic classifiers.

1. Introduction and Motivation

Many graphical representations and tools for classifier evaluation have been proposed in the literature, such as ROC curves and isometrics (Swets et al., 2000; Flach, 2003; Fawcett, 2006), DET curves (Martin et al., 1997), lift charts (Piatetsky-Shapiro & Masand, 1999), calibration maps (Cohen & Goldszmidt, 2004), among many others. Some of these visualise two performance metrics as a function of an implicit operating condition: e.g., ROC curves, precision-recall curves and DET curves. Others have the operating condition explicitly on the *x*-axis, and a single performance metric (accuracy, error rate, loss) on the *y*-axis. Cost curves (Drummond & Holte, 2000; 2006) are in the latter category. The basic idea is to draw loss on the *y*-axis against 'probability cost' (an operating condition depending on both class and misclassification cost distribution, called skew in this paper) on the *x*-axis. For a fixed decision threshold there is a linear relationship between skew and loss; this is called a cost line. The lowest cost line for a given skew then represents the optimal loss achievable with the classifier for that skew, and the cost curve is the lower envelope of the cost lines.

There are several correspondences between ROC space and cost space, the most important of which is a point-line duality: line segments in ROC space correspond to points in cost space and points in ROC space correspond to line segments in cost space. Furthermore, the convex hull of a ROC curve corresponds to the lower envelope of the cost lines in cost space. There are also differences, arising from the fact that ROC curves concentrate on ranking performance while cost curves visualise classification performance. Furthermore, (Drummond & Holte, 2006) did not propose a cost curve equivalent of a non-convex ROC curve. In other words, cost curves represent the performance of the ROC convex hull of a classifier, which is a typically optimistic (and frequently unrealistic) assessment of a classifier. Another correspondence between ROC curves and cost curves is that they both ignore the magnitude of the scores. As a result, neither the ROC curve nor the cost curve are affected by a monotonic transformation of the scores assigned by a classifier. This is natural for ROC curves, as they are designed for evaluating ranking performance. It appears to be

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

less natural for cost curves, as without knowing anything about the scale of the classifier's scores it is impossible to set decision thresholds in a uniform, classifier-independent way. We conclude, then, that it is this reluctance to take score magnitudes into account that necessitates the overly optimistic and unrealistic assumption of optimal threshold selection.

The alternative proposed in this paper is to assume that the classifier scores are posterior class probabilities, which gives us a natural way of choosing the thresholds. Namely, given an operating condition o, we will predict positive if the score is greater than o and negative otherwise. With this simple decision rule, we can visualise the loss for arbitrary operating conditions in cost space, which produces a new curve, also spanned by cost lines, and hence never below the cost curve. This new curve clearly depends on the quality of the probability estimates, and it shows the performance for the full range of operating conditions. This implies that we can work in a way similar to ROC analysis: we can choose and discard classifiers depending on the operating conditions but we can also combine classifiers in order to obtain a lower overall loss. Furthermore, the area under the new curve has a natural interpretation as the Brier score or mean squared error of the probability estimates, which justifies the name we have given.

The outline of the paper is as follows. Section 2 formally introduces cost curves and the necessary notation for the rest of the paper. Section 3 considers a more natural way of calculating the expected loss, which produces the new curves. The main result, that the area under the new curve is the Brier score, is proved in Section 4. Section 5 shows how Brier curves can be used to compare probabilistic classifiers and to derive combined classifiers in order to improve the joint Brier curve. Finally, Section 6 concludes with a short discussion and possible areas for further work.

2. Preliminaries

In this section we present our notational conventions and discuss previous work on cost curves. (Much of this section is shared with a related paper investigating the properties of *AUC* as a measure of aggregated classification performance (Flach et al., 2011).)

2.1. Notation

The instance space is denoted X and the output space Y. Elements in X and Y will be referred to as x and y respectively. For this paper we will assume binary classifiers, i.e., $Y = \{0,1\}$. A crisp or categorical classifier is a function that maps examples to classes. A probabilistic classifier 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. A scoring classifier is a function $m : X \to \Re$ that maps examples to real numbers on an unspecified scale, such that scores are monotonically related to $\hat{p}(1|x)$. In order to make predictions in the Y domain, a probabilistic or scoring classifier can be converted to a crisp 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 classifier 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; $1 - F_1(t)$ is the true positive rate or sensitivity.[1]

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 $n_k = |D_k|$ and $\pi_k = n_k/n$. We will use the term *class proportion* for π_0 (other terms such as 'class ratio' or 'class prior' have been used in the literature). Given any given strict order for a dataset of n examples we will use the index i on that order to refer to the i-th example. Thus, s_i denotes the score of the i-th example and y_i its true class. Given a dataset and a classifier, we can define empirical score distributions for which we will use the same symbols as the population functions. We then have $f_k(s) = \frac{1}{n_k}|\{\langle x,y \rangle \in D_k | m(x) = s\}|$ and $F_k(t) = \sum_{s \leq t} f_k(s)$.

The Brier score is a well-known evaluation measure for probabilistic classifiers. It is an alternative name for the Mean Squared Error or MSE loss (Brier, 1950). $BS(m,D)$ is the Brier score of classifier m with data D; we will usually omit m and D when clear from the context. We define $BS_k(m,D) = BS(m,D_k)$. The Brier score is defined as $BS \triangleq \frac{1}{n}\sum_{i=1}^{n}(s_i - y_i)^2$, where s_i is the score predicted for example i and y_i is the true class for example i. Clearly, $BS = \pi_0 BS_0 + \pi_1 BS_1$. The corresponding population quantities are $BS_0 = \int_0^1 s^2 f_0(s)ds$ and $BS_1 = \int_0^1 (1-s)^2 f_1(s)ds$.

2.2. Loss, ROC Curves and Cost Curves

An operating condition or deployment context is usually defined by a class distribution and a way to aggregate misclassification cost over examples. One general approach to cost-sensitive learning assumes that the cost does not depend on the example but only on its class. In this way, mis-

[1] We use 0 for the positive class and 1 for the negative class, but scores increase with $\hat{p}(1|x)$. That is, a ranking from strongest positive prediction to strongest negative prediction has non-decreasing scores. This is the same convention as used by, e.g., (Hand, 2009).

classification costs are usually simplified by means of cost matrices, where we can express that some misclassification costs are higher than others (Elkan, 2001). Typically, the costs of correct classifications are assumed to be 0. This means that for binary classifiers we can describe the cost matrix by two values $c_k \geq 0$, representing the misclassification cost of an example of class k. Additionally, we can normalise the costs by setting $b = c_0 + c_1$ and $c = c_0/b$; we will refer to c as the *cost proportion*. Since b is a constant which only affects the magnitude of the costs but is independent of the classifier, we will set $b = 2$ which has the advantage that loss is commensurate with error rate which assumes $c_0 = c_1 = 1$.

The loss which is produced at a decision threshold t and a cost proportion c is then given by the formula:

$$\begin{aligned} Q_c(t;c) &\triangleq c_0\pi_0(1-F_0(t)) + c_1\pi_1 F_1(t) \\ &= 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\} \end{aligned} \tag{1}$$

We often are interested in analysing the influence of class proportion and cost proportion at the same time. Since the relevance of c_0 increases with π_0, an appropriate way to consider both at the same time is by the definition of *skew*, which is a normalisation of their product:

$$z \triangleq \frac{c_0\pi_0}{c_0\pi_0 + c_1\pi_1} = \frac{c\pi_0}{c\pi_0 + (1-c)(1-\pi_0)} \tag{2}$$

From Eq. (1) we obtain

$$\frac{Q_c(t;c)}{c_0\pi_0 + c_1\pi_1} = z(1-F_0(t)) + (1-z)F_1(t) \triangleq Q_z(t;z) \tag{3}$$

This gives an expression for loss at a threshold t and a skew z. We will assume that the operating condition is either defined by the cost proportion (using a fixed class distribution) or by the skew.

The ROC curve (Swets et al., 2000; Fawcett, 2006) is defined as a plot of $F_1(t)$ (i.e., false positive rate at decision threshold t) on the x-axis against $F_0(t)$ (true positive rate at t) on the y-axis, with both quantities monotonically non-decreasing with increasing t (remember that scores increase with $\hat{p}(1|x)$ and 1 stands for the negative class). We then have that the Area Under the ROC curve (AUC) can be defined as

$$AUC = \int_0^1 F_0(s)dF_1(s) = \int_{-\infty}^{+\infty} F_0(s)f_1(s)ds \tag{4}$$

When dealing with empirical distributions the integral is replaced by a sum.

The convex hull of a ROC curve (ROCCH) is a construction over the ROC curve in such a way that all the points on the ROCCH have minimum loss for some choice of c or z. This means that we restrict attention to the *optimal* threshold for a given cost proportion c:

$$\begin{aligned} T_c^o(c) &\triangleq \arg\min_t \{Q_c(t;c)\} \\ &= \arg\min_t \{2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\} \end{aligned} \tag{5}$$

which matches the optimal threshold for a given skew z:

$$T_z^o(z) \triangleq \arg\min_t \{Q_z(t;z)\} = T_c^o(c) \tag{6}$$

The convex hull is defined by linear interpolation between the points $\{F_1(t), F_0(t)\}$ where $t = T_c^o(c)$ for some c. The Area Under the ROCCH (denoted by $AUCH$) can be computed in a similar way as the AUC with modified versions of f_k and F_k. Obviously, $AUCH \geq AUC$, with equality implying the ROC curve is convex.

A cost plot as defined by (Drummond & Holte, 2006) has $Q_z(t;z)$ on the y-axis against skew z on the x-axis (Drummond and Holte use the term 'probability cost' rather than skew). Since $Q_z(t;z) = z(1-F_0(t)) + (1-z)F_1(t)$, cost lines for a given decision threshold t are straight lines $Q_z = a + bz$ with intercept $a = F_1(t)$ and slope $b = 1 - F_0(t) - F_1(t)$. A cost line visualises how cost at that threshold changes between $F_1(t)$ for $z = 0$ and $1 - F_0(t)$ for $z = 1$. The cost curve is then the lower envelope of all the cost lines, obtained by only considering the optimal threshold (the lowest cost line) for each skew. An explicit definition of the cost curve as a function of z in our notation is

$$CC(z) \triangleq Q_z(T_z^o(z);z) \tag{7}$$

Example 1. *Figure 1 (left) shows a ROC curve and a cost curve (right) for a classifier with 4 examples of class 1 and 11 examples of class 0. Because of ties, there are 11 distinct scores. We observe 7 segments in the original ROC curve on the left, and 5 segments in its convex hull. We see that these 5 segments correspond to the 5 points in the cost curve on the right. The cost curve is 'constructed' as the lower envelope of the 12 cost lines (one more than the number of distinct scores). The middle plot is an alternative cost plot with cost proportion rather than skew on the x-axis. That is, here the cost lines are straight lines $Q_c = a' + b'c$ with intercept $a' = 2\pi_1 F_1(t)$ and slope $b' = 2\pi_0(1-F_0(t)) - 2\pi_1 F_1(t)$. We can clearly observe the class imbalance.*

We see a clear correspondence between the ROC convex hull in ROC space and the cost curve in cost space, but no cost space equivalent of the non-convex ROC curve. We propose such an equivalent in the next section.

3. Brier Curves

As shown above, cost curves assume that we set thresholds optimally, choosing the same thresholds as the ROC convex hull according to Eqs. (5-6). However, thresholds that

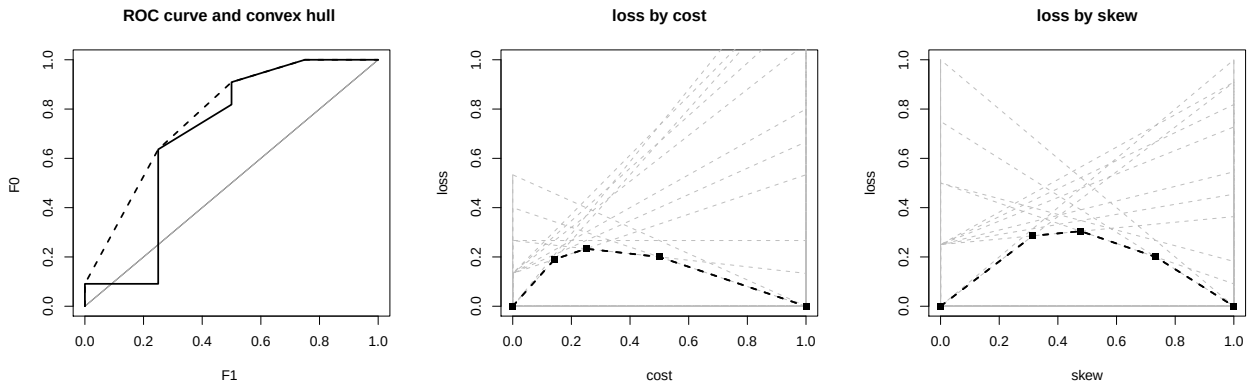

Figure 1. Several graphical representations for a classifier with probability estimates (0.05 0.15 0.16 0.18 0.20 0.20 0.45 0.55 0.70 0.70 0.70 0.85 0.90 0.90 0.95) and classes (0 1 0 0 0 0 0 0 0 1 0 0 1 0 1). Left: ROC curve and convex hull. Middle: cost lines and cost curve against cost proportions. Right: cost lines and cost curve against skews.

are optimal on a validation set may not carry over to a new test set. In other words, assuming optimal thresholds leads to an optimistic assessment of performance. A more natural way of setting the threshold for a probabilistic classifier is to consider its probability estimates. We simply set the threshold to the operating condition, either cost proportion or skew. This takes the magnitudes of the scores into account, which is desirable as we argued in the introduction.

Definition 1 (Probabilistic threshold choice)**.** *For a given probabilistic classifier and operating condition defined by cost proportion, the probabilistic threshold choice method sets the threshold as follows:*

$$T_c^p(c) \triangleq c \tag{8}$$

If the operating condition is defined by skew, the threshold is set as follows:

$$T_z^p(z) \triangleq z = T_c^p(c)\frac{z}{c} \tag{9}$$

We can use this probabilistic threshold choice method to define a new kind of curve in cost space.

Definition 2 (Brier curve)**.** *The* Brier curve *for a given classifier is defined as a plot of loss against operating condition using the probabilistic threshold choice method. In particular, if the operating condition is determined by cost proportion the Brier curve is defined by*

$$\begin{aligned} BC_c(c) &\triangleq Q_c(T_c^p(c);c) = Q_c(c;c) \\ &= 2c\pi_0(1-F_0(c)) + 2(1-c)\pi_1 F_1(c) \end{aligned} \tag{10}$$

A Brier curve for skew is defined by

$$\begin{aligned} BC_z(z) &\triangleq Q_z(T_z^p(z);z) = Q_z(z;z) \\ &= z(1-F_0(z)) + (1-z)F_1(z) \end{aligned} \tag{11}$$

We will drop the subscript indicating the type of operating condition if it is clear from the context. We will sometimes decompose the curve by class: e.g., for cost proportions $BC_0(c) = 2c\pi_0(1-F_0(c))$ *and* $BC_1(c) = 2(1-c)\pi_1 F_1(c)$.

We first concentrate on the Brier curve defined in terms of cost proportions (we will justify the name in the next section). Like the cost curve it is piecewise linear as it consists of segments of cost lines (that is, if we work with empirical distributions). However, unlike the cost curve it has discontinuities at points where the cost proportion (and hence the threshold) equals the score of one of the examples. These discontinuities mirror the discontinuities in the (empirical) cumulative distribution functions.[2] In contrast, cost curves are continuous because, as a result of choosing thresholds optimally, they are the lower envelope of all the cost lines. Both curves have discontinuities in their first derivative.

Example 2. *Figure 2 shows the Brier curve of the classifier from Example 1. The curve is piecewise linear with discontinuities at points where the cost proportion c equals the score of one of the examples: at these points the Brier curve 'jumps' to one of the other cost lines. For example, the biggest jump occurs at* $c = 0.7$*, as this threshold changes the classification of two positives and one negative and hence has a big influence on overall loss.*

We can also see that the highest loss is obtained for a cost proportion just below 0.45*. The cost curve peaks at a much lower value of* $c = 0.25$*. More generally, the cost curve is everywhere below the Brier curve except for* $c < 0.15$ *and* $c \geq 0.90$*. The Brier curve therefore exactly quantifies the extent to which the cost curve is optimistic, and over what operating range.*

The Brier curve also suggests changes that could be made to the scores in order to improve the curve. For instance, if the examples with score 0.70 *are changed to* 0.60*, this would clearly lower the Brier curve in that interval.*

[2] Like empirical cumulative distribution functions, Brier curves are right continuous with left limits.

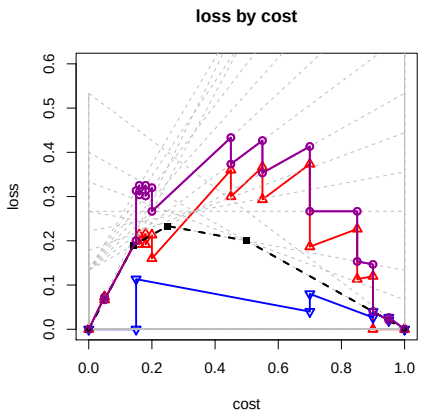

Figure 2. The top curve is the Brier curve of the classifier from Example 1, depicting the loss if the decision threshold is set equal to the cost proportion c. The two discontinuous curves below are BC_0 and BC_1, respectively. The cost curve is shown as a thick dashed line. We can see that the probabilistic threshold choice method is suboptimal between $c = 0.15$ and $c = 0.90$.

As a more realistic example, Figure 3 shows the Brier curves of a J48 model trained in Weka (Witten & Frank, 2005) on the credit rating dataset from the UCI repository (Frank & Asuncion, 2010) with a 50%-50% train-test split. The classifier on the top plots is J48 with default parameters (pruning enabled, Laplace correction disabled), while the bottom classifier is J48 without pruning but with Laplace smoothing. We can clearly see the overfitting of the unpruned tree, as it shows considerable difference between the (good) training set curve and the (bad) test set curve. We can also see the effect of the Laplace correction, which deliberately sacrifices training set performance on extreme cost proportions in the hope of better generalisation performance. On the test set, we see that estimated probabilities are well-calibrated for high cost proportions but not for low ones.

4. The Area under the Brier Curve is the Brier Score

Since the Brier curve plots loss against operating condition, the area under it is expected loss, averaged over the whole operating range. Let us concentrate first on cost proportion as operating condition. The expected loss is defined as

$$L_c \triangleq \int_0^1 BC_c(c)dc = \int_0^1 Q_c(c;c)dc$$
$$= \int_0^1 2\{c\pi_0(1-F_0(c)) + (1-c)\pi_1 F_1(c)\}dc \qquad (12)$$

We then have the following result.

Theorem 1. *The area under the Brier curve for cost proportions is equal to the Brier score.*

Proof. We have $BS = \pi_0 BS_0 + \pi_1 BS_1$. Using integration by parts, we have

$$BS_0 = \int_0^1 s^2 f_0(s)ds = \left[s^2 F_0(s)\right]_{s=0}^1 - \int_0^1 2sF_0(s)ds$$
$$= 1 - \int_0^1 2sF_0(s)ds = \int_0^1 2sds - \int_0^1 2sF_0(s)ds$$

Similarly for the negative class:

$$BS_1 = \int_0^1 (1-s)^2 f_1(s)ds$$
$$= \left[(1-s)^2 F_1(s)\right]_{s=0}^1 + \int_0^1 2(1-s)F_1(s)ds$$
$$= \int_0^1 2(1-s)F_1(s)ds$$

Taking their weighted average, we obtain

$$BS = \pi_0 BS_0 + \pi_1 BS_1$$
$$= \int_0^1 \{\pi_0(2s - 2sF_0(s)) + \pi_1 2(1-s)F_1(s)\}ds$$

which, after reordering of terms and change of variable, is the same expression as Eq. (12). □

The proof for the empirical case, where the cumulative distribution functions F_0 and F_1 are piecewise constant and discontinuous, is similar but more involved notationally.

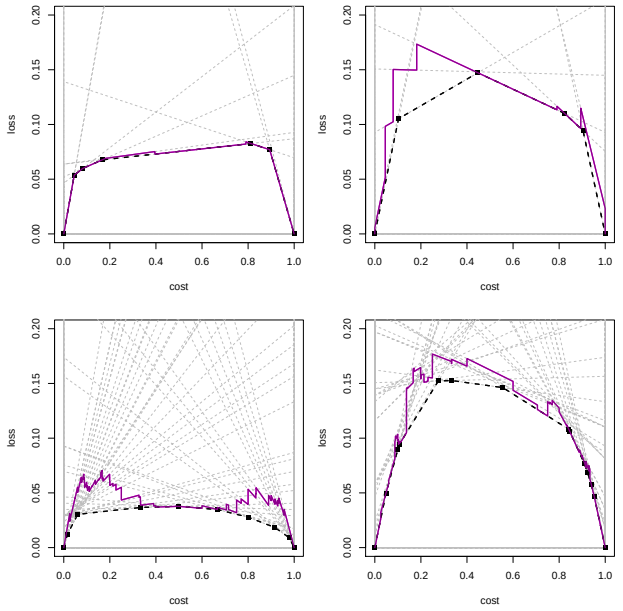

Figure 3. Brier curves and cost curves for two different J48 classifiers evaluated on training and test sets both sampled from the credit rating UCI dataset. Top left: Pruned tree on training set (*AUC*: 0.937, *AUCH*: 0.937, *BS*: 0.068). Top right: Pruned tree on test set (*AUC*: 0.887, *AUCH*: 0.894, *BS*: 0.126). Bottom left: Unpruned tree on training set (*AUC*: 0.985, *AUCH*: 0.988, *BS*: 0.042). Bottom right: Unpruned tree on test set (*AUC*: 0.893, *AUCH*: 0.904, *BS*: 0.126).

We believe that this connection between Brier curves and the Brier score leads to a better understanding of both concepts. The fact that the area under the Brier curve has an independent meaning in terms of a well-known performance index lends further credibility to Brier curves[3] similar to the way that the interpretation of *AUC* as the Wilcoxon-Mann-Whitney sum of ranks statistic lends credibility to ROC curves. Conversely, Brier curves offer a generalisation of the Brier score in the sense that we can investigate 'partial Brier score' as expected loss over a more restricted range of operating conditions, and ultimately test the contribution of the score differences between individual examples. In other words, the Brier curve can be seen as an example-wise decomposition of the Brier score, quite different from the well-known decomposition in terms of calibration and refinement (Murphy, 1973).

For completeness we state the corresponding result for skews. We define expected loss as

$$L_z \triangleq \int_0^1 BC_z(z)dz = \int_0^1 Q_z(z;z)dz$$
$$= \int_0^1 \{z(1-F_0(z)) + (1-z)F_1(z)\}dz \qquad (13)$$

Corollary 2. $L_z = (BS_0 + BS_1)/2$.

5. Comparing Classifiers and Building Hybrid Classifiers using Brier Curves

One of the most useful features of ROC analysis is that we can compare classifiers and identify regions where one classifier dominates other classifiers. This makes it possible to choose operating ranges and to discard classifiers safely. However, it should be noted that neither operating condition nor decision thresholds are explicitly represented in ROC plots. Cost curves have the operating condition on the x-axis but no representation of corresponding optimal thresholds. With Brier curves we assume that thresholds are chosen using the probability estimates from the classifier, which is exactly what they are for. So, given an operating condition on the x-axis we can simply read off on the y-axis which classifier will have lowest loss. Given two classifiers A and B we say that A *dominates* B at a cost proportion c iff $Q_c^A(c;c) < Q_c^B(c;c)$. From here we can define dominance intervals and even discard classifiers completely if they do not dominate in any interval.

Example 3. *Consider the following scores and ranks (between parentheses) assigned by three classifiers A, B and C to a dataset consisting of 4 negatives and 6 positives:*

[3] In (Murphy, 1966) we find a similar relation to expected utility (in our notation, $-(1/4)PS + (1/2)(1+\pi_0)$), where the so-called probability score is $PS = 2BS$. The differences arise because Murphy works with utilities rather than costs and uses a different cost matrix.

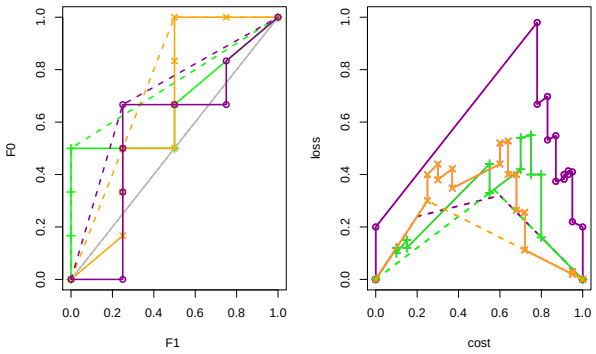

Figure 4. ROC curves and convex hulls (left) and Brier curves and cost curves (right) for three of the classifiers in Example 3: green lines with '+' points: classifier A (*AUC*: 0.667, *AUCH*: 0.750, *BS*: 0.244); orange lines with 'x' points: classifier B (*AUC*: 0.646, *AUCH*: 0.750, *BS*: 0.240); magenta lines with 'o' points: classifier C (*AUC*: 0.563, *AUCH*: 0.708, *BS*: 0.558).

	Class	*A*	*B*	*C*	*D*
e_1	1	0.70 (4..5)	0.60 (5)	0.00 (1)	0.65 (5)
e_2	1	0.80 (7..10)	1.00 (10)	1.00 (9..10)	0.90 (10)
e_3	1	0.80 (7..10)	0.95 (9)	0.93 (7)	0.88 (9)
e_4	1	0.70 (4..5)	0.25 (1..2)	0.91 (6)	0.48 (4)
e_5	0	0.80 (7..10)	0.68 (7)	0.78 (2..3)	0.74 (7)
e_6	0	0.75 (6)	0.64 (6)	0.83 (4)	0.70 (6)
e_7	0	0.10 (1)	0.37 (4)	0.78 (2..3)	0.24 (2)
e_8	0	0.55 (3)	0.30 (3)	0.95 (8)	0.43 (3)
e_9	0	0.80 (7..10)	0.72 (8)	1.00 (9..10)	0.76 (8)
e_{10}	0	0.15 (2)	0.25 (1..2)	0.87 (5)	0.20 (1)

Figure 4 shows ROC curves, ROC convex hulls, Brier curves and cost curves for classifiers A, B and C. In terms of the Brier curve, classifier A dominates classifier B from 0.1 to 0.5 and from 0.55 to 0.64, while B dominates A from 0.5 to 0.55 and from 0.64 to 1, and neither dominates the other from 0 to 0.1. However, C is dominated by both A and B in the entire cost proportion range. So, classifier C can be safely discarded if thresholds are chosen in a probabilistic way. Notice that the poor performance of C is to a large part caused by the large difference in predicted probability of the first and second example in the ranking, which is very clearly visualised by the Brier curve.

In terms of cost curves, these dominance regions are different: Classifier A dominates classifier B from 0 to 0.4 and B dominates A from 0.4 to 1. Furthermore, there is a small operating range where C dominates A, and another one where C dominates B. Only by using the convex hull of A and B can we discard C completely. This can also be seen in the ROC plots.

We conclude from this example that, not only are the dominance regions different, but there are also cases where a classifier can be discarded using one kind of threshold choice method but not using the other. It is also possible to construct examples where we can discard a classifier using ROC analysis but not by means of Brier curves.

Related to the notion of dominance is the idea of combining classifiers, or modifying a classifier in a given operating

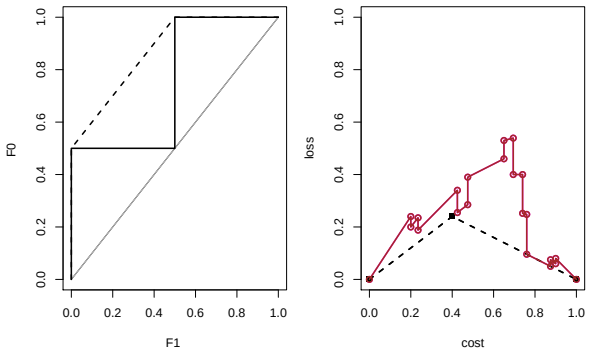

Figure 5. ROC curve and Brier curve of classifier *D* which predicts the average of the probabilities predicted by classifiers *A* and *B* in Example 3 (*AUC*: 0.750, *AUCH*: 0.875, *BS*: 0.231).

range, in order to improve performance. This is a well-known procedure in ROC space. For instance, concavities in the ROC curve of a scoring classifier can be repaired by randomising or inverting the ranking in the corresponding operating range (Flach & Wu, 2005). The latter procedure can also be carried out in cost space as it involves taking the lower envelope of the cost lines involved in a particular operating range of the Brier curve.

Brier curves open up new ways of combining classifiers on the basis of their Brier curves. Three possibilities present themselves. The first is to make a random choice between two probabilistic classifiers for each prediction,. This has the effect of averaging the two Brier curves and the corresponding Brier scores, and may help to obtain a more robust classifier. The second possibility is to average the predicted probabilities of the classifiers – which is not the same thing. Figure 5 shows the Brier curve of classifier *D* in Example 3, which predicts the average of classifiers *A* and *B*. As we see, the resulting classifier is slightly better than *A* and *B* alone in terms of *AUC*, *AUCH* and *BS*.

However, averaging the scores does not get the best of *A* and *B*. From the Brier curves we can construct a hybrid classifier *AB*, which uses *A*'s predictions if the cost proportion is in either interval $[0.1, 0.5]$ or $[0.55, 0.65]$ and *B*'s predictions otherwise. This hybrid classifier is a meta-classifier that cannot be represented by a single set of scores. Hybrid classifiers which are piecewise constructed from sections are biased, as (Drummond & Holte, 2006) recognise: "a hybrid classifier built piecewise from the cost curves for the individual classifiers that make up the hybrid is not an unbiased estimate of performance except when a performance-independent selection criterion is used" (Section 5.2). The way in which piecewise constructed classifiers using Brier curves may be more or less biased depend on the quality of the probability estimates, i.e., on the extent to which they are calibrated.

In Figure 6 we demonstrate the impact of calibration using Brier curves. On the left we see the ROC curves and Brier

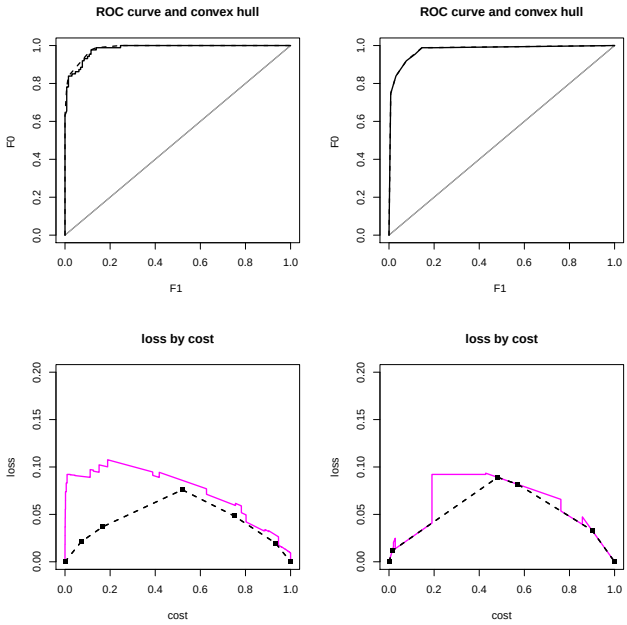

Figure 6. ROC curves and Brier curves for a Naive Bayes classifier on the vote UCI dataset before and after PAV calibration. Top left: Non-calibrated ROC curve. Top right: PAV-calibrated ROC curve. Bottom left: Non-calibrated Brier curve. Bottom right: PAV-calibrated Brier curve.

curves obtained from the raw probabilities of Naive Bayes using the vote UCI dataset (50% train, 50% test). The Brier curve clearly locates the loss due to bad calibration between scores 0 and 0.5, although this has little effect on the ranking quality. On the right we see the result of calibrating probabilities on the training data with the PAV algorithm (Fawcett & Niculescu-Mizil, 2007). As expected, calibration improves both curves. With ROC curves, calibration has the potential to fix the concavities of the curve, while with Brier curves it moves the curve closer to the optimal cost curve. We can see clearly that calibration has failed between 0.2 and 0.4, which corresponds to the strong discontinuity of the slope of the ROC curve.

6. Concluding Remarks

In this paper we have introduced a new graphical tool to understand the performance of classifiers. This tool is a curve, drawn in cost space, which allows us to see the performance of a probabilistic classifier for a range of operating conditions defined by cost proportion or skew. While ROC curves are useful to represent and analyse rankers, Brier curves are useful to represent probabilistic classifiers. In fact, we can operate with Brier curves in a similar way to ROC curves through the notion of dominance.

Like the ROC convex hull, cost curves take the overly optimistic view that thresholds are chosen optimally to min-

imise cost. Analysing classifier performance using Brier curves is more appropriate when the probability estimates are used to set the thresholds, which is both more common and more realistic. Additionally, the difference in cost space area between two classifiers is measured as a proportion of their Brier scores, which means that we can detect the areas where each classifier increases its Brier score over others, or cases where two classifiers have similar Brier score but different Brier curves (such as classifiers *A* and *B* in Figure 4). Jointly, cost curves and Brier curves summarise most of the information about the performance of a classifier, and allow to consider different ways of choosing the thresholds, and their resulting performance. Consequently, we think that cost curves and Brier curves are perfect companions.

Brier curves, and their connection to the Brier score, open up many interesting lines to pursue. For instance, confidence intervals for the Brier curve and confidence intervals for the Brier score are expected to be related. Brier curves can also play a role in the improvement of classifiers, especially in terms of calibration. For instance, the Brier score decomposition and the notion of calibration is likely to have an interpretation in cost space. Finally, the various ways in which Brier curves can be combined to build hybrid classifiers is interesting to explore.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. This work has been partially supported by the EU (FEDER) and the Spanish MICINN, under grant TIN2010-21062-C02-02, the Spanish project 'Agreement Technologies' (Consolider Ingenio CSD2007-00022) and the GVA project PROMETEO/2008/051.

References

Brier, G.W. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Cohen, I. and Goldszmidt, M. Properties and benefits of calibrated classifiers. *Knowledge Discovery in Databases: PKDD 2004*, pp. 125–136, 2004.

Drummond, C. and Holte, R.C. Explicitly representing expected cost: An alternative to ROC representation. In *Knowl. Discovery and Data Mining*, pp. 198–207, 2000.

Drummond, C. and Holte, R.C. Cost Curves: An Improved Method for Visualizing Classifier Performance. *Machine Learning*, 65(1):95–130, 2006.

Elkan, C. The foundations of Cost-Sensitive learning. In Nebel, Bernhard (ed.), *Proceedings of the seventeenth International Conference on Artificial Intelligence (IJCAI-01)*, pp. 973–978, 2001.

Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Fawcett, T. and Niculescu-Mizil, A. PAV and the ROC convex hull. *Machine Learning*, 68(1):97–106, 2007.

Flach, P.A. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, pp. 194–201, 2003.

Flach, P.A. and Wu, S. Repairing concavities in ROC curves. In Kaelbling, L.P. and Saffiotti, A. (eds.), *Proceedings of the 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05)*, pp. 702–707, 2005.

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

Frank, A. and Asuncion, A. UCI machine learning repository, 2010.

Hand, D.J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1):103–123, 2009. ISSN 0885-6125.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. The DET curve in assessment of detection task performance. In *Fifth European Conference on Speech Communication and Technology*. Citeseer, 1997.

Murphy, A.H. A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *Journal of Applied Meteorology*, 5: 534–536, 1966. ISSN 0894-8763.

Murphy, A.H. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, 12:595–600, 1973. ISSN 0894-8763.

Piatetsky-Shapiro, G. and Masand, B. Estimating campaign benefits and modeling lift. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 193. ACM, 1999.

Swets, J.A., Dawes, R.M., and Monahan, J. Better decisions through science. *Scientific American*, 283(4):82–87, October 2000.

Witten, I.H. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.