Explicitly Representing Expected Cost: An Alternative to ROC Representation

Chris Drummond
School of Information Technology and Engineering
University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5
cdrummon@site.uottawa.ca

Robert C. Holte
School of Information Technology and Engineering
University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5
holte@site.uottawa.ca

ABSTRACT
This paper proposes an alternative to ROC representation, in which the expected cost of a classifier is represented explicitly. This expected cost representation maintains many of the advantages of ROC representation, but is easier to understand. It allows the experimenter to immediately see the range of costs and class frequencies where a particular classifier is the best and quantitatively how much better it is than other classifiers. This paper demonstrates there is a point/line duality between the two representations. A point in ROC space representing a classifier becomes a line segment spanning the full range of costs and class frequencies. This duality produces equivalent operations in the two spaces, allowing most techniques used in ROC analysis to be readily reproduced in the cost space.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning,Induction*

General Terms
ROC Analysis, Cost Sensitive Learning

1. INTRODUCTION
Provost and Fawcett [9] have argued persuasively that accuracy is often not an appropriate measure of classifier performance. This is certainly apparent in classification problems with heavily imbalanced classes (one class occurs much more often than the other). It is also apparent when there are asymmetric misclassification costs (the cost of misclassifying an example from one class is much larger than the cost of misclassifying an example from the other class). Class imbalance and asymmetric misclassification costs are related to one another. One way to correct for imbalance is to train a cost sensitive classifier with the misclassification cost of the minority class greater than that of the majority class, and one way to make an algorithm cost sensitive is to intentionally imbalance the training set. As an alternative to accuracy, Provost and Fawcett advocate the use of ROC analysis, which measures classifier performance over the full range of possible costs and class frequencies. They also proposed the convex hull as a way to determine the best classifier for a particular combination of costs and class frequencies.

Decision theory can be used to select the best classifier if the costs and class frequencies are known ahead of time. But often they are not fixed until the time of application making ROC analysis important. The relationship between decision theory and ROC analysis is discussed in Lusted's book [7]. In Fawcett and Provost's [4, 5] work on cellular fraud detection, they noted that the cost and amount of fraud varies over time and location. This was one motivation for their research into ROC analysis. Our own experience with imbalanced classes [6] dealt with the detection of oil spills and the number of non-spills far outweighed the number of spills. Not only were the classes imbalanced, the distribution of spills versus non-spills in our experimental batches was unlikely to be the one arising in practice. We also felt that the trade-off between detecting spills and false alarms was better left to each end user of the system. These considerations led to our adoption of ROC analysis. Asymmetric misclassification costs and highly imbalanced classes often arise in Knowledge Discovery and Data Mining (KDD) and Machine Learning (ML) and therefore ROC analysis is a valuable tool in these communities.

In this paper we focus on the use of ROC analysis for the visual analysis of results during experimentation, and the interactive KDD process, and the presentation of those results in reports. For this purpose despite all of the strengths of the ROC representation, we found the graphs produced were not always easy to interpret. Although it is easy to see which curve is better in figure 1, it is much harder to determine by how much. It is also not immediately clear for what costs and class distributions classifier A is better than classifier B. Nor is it easy to "read-off" the expected cost of a classifier for a fixed cost and class distribution. In figure 2 one curve is better than the other for some costs and class distributions, but the range is not determined by the crossover point of the curves so is not immediately obvious. This information can be extracted as it is implicit in the graph, but our alternative representation makes it explicit.

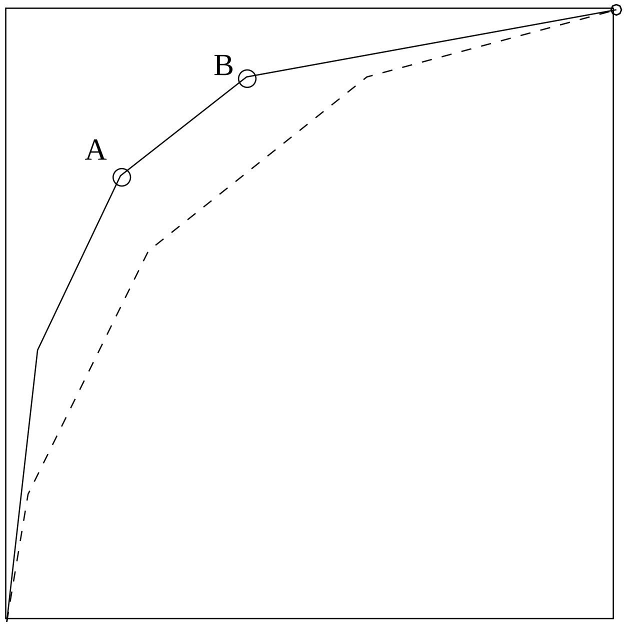

Figure 1: Comparing Performance

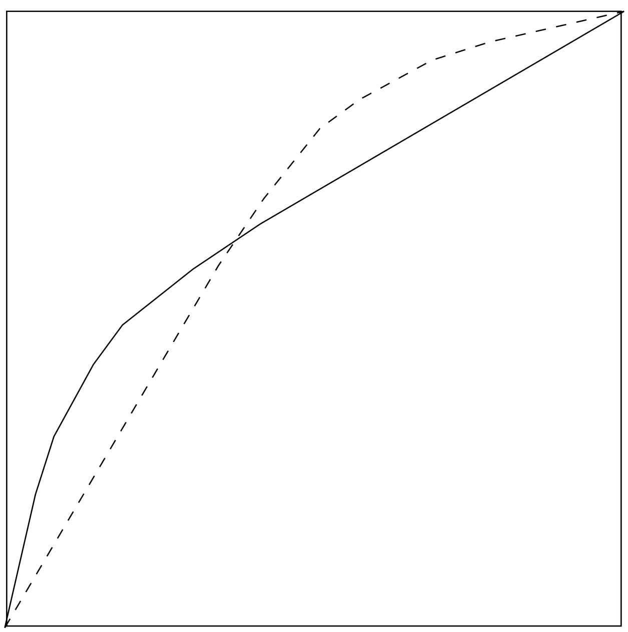

Figure 2: Performance Ranges

2. TWO DUAL REPRESENTATIONS

In this section we briefly review ROC analysis and how it is used in evaluating or comparing a classifier's performance. We then introduce our alternative dual representation, which maintains these advantages but by making explicit the expected cost is much easier to understand. In both representations, the analysis is restricted to two class problems which are referred to as the positive and negative class.

2.1 The ROC Representation

Provost and Fawcett [9] are largely responsible for introducing ROC analysis to the KDD and ML communities. It had been used extensively in signal detection, where it earned its name "Receiver Operating Characteristics" abbreviated to ROC. Swets [12] showed that it had a much broader applicability, by demonstrating its advantages in evaluating diagnostic systems. In ROC analysis instead of just a single value of accuracy, a pair of values is recorded for different costs and class frequencies. In signal detection these were called the hit rate and false alarm rate. In the KDD and ML communities they are called the true positive rate (the fraction of positives correctly classified) and false positive rate (the fraction of negatives misclassified). This pair of values produces a point in ROC space: the false positive rate being the x-coordinate, the true positive rate being the y-coordinate.

Some classifiers have parameters for which different settings produce different ROC points. For example, a classifier that produces probabilities of an example being in each class, such as a Naive Bayes classifier, can have a threshold parameter biasing the final class selection [3, 8]. Plotting all the ROC points that can be produced by varying these parameters produces an ROC curve for the classifier. Typically this is a discrete set of points, including (0,0) and (1,1), which are connected by line segments. If such a parameter does not exist, algorithms such as decision trees can be modified to include costs to produce the different points [2]. Alternatively the class frequencies in the training set can be changed by under or over sampling to simulate a change in class priors or misclassification costs [3].

One point in an ROC diagram dominates another if it is above and to the left, i.e. has a higher true positive rate (TP) and a lower false positive rate (FP). If point A dominates point B, A it will have a lower expected cost than B for all possible cost ratios and class distributions. One set of points A is dominated by another B when each point in A is dominated by some point B and no point in B is dominated by a point in A. The normal assumption in ROC analysis is that these points are samples of a continuous curve and therefore normal curve fitting techniques can be used. In Swets's work [12] smooth curves are fitted to typically a small number of points, say four or five. Alternatively a non-parametric approach is to use a piece-wise linear function, joining adjacent points by straight lines. Dominance is then defined for all points on the curve.

Traditional ROC analysis has as its primary focus determining which diagnostic system or classifier has the best performance independent of cost or class frequency. But there is also an important secondary role of selecting the set of system parameters (or individual classifier) that gives the best performance for a particular cost or class frequency. This can be done by means of the upper convex hull of the points, which has been shown to dominate all points under the hull [9]. It has further been shown that dominance implies superior performance for a variety of commonly-used performance measures [10]. The dashed line in figure 3 is a typical ROC convex hull. The slope of a segment of the convex hull connecting the two vertices (FP_1, TP_1) and (FP_2, TP_2) is given by equation 1.

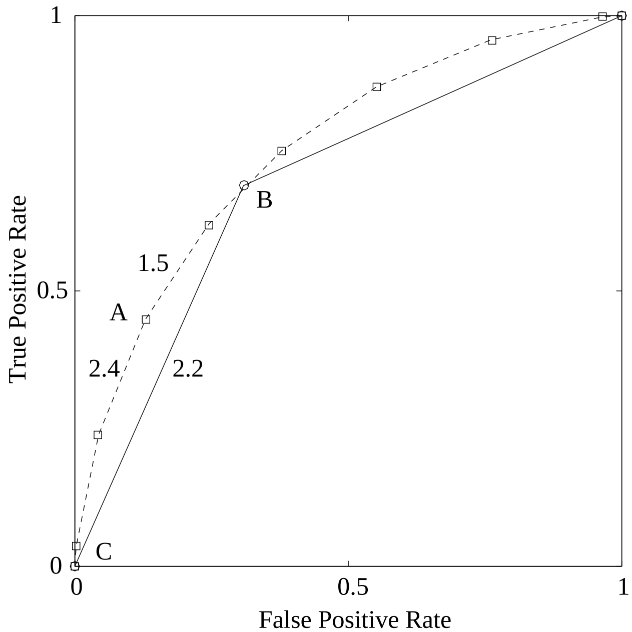

Figure 3: Comparing Two ROC curves

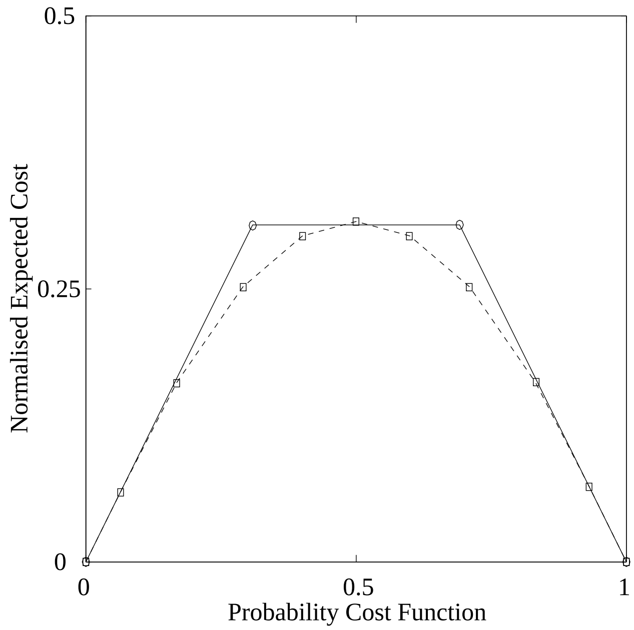

Figure 4: Comparing Misclassification Costs

$$\frac{TP_1 - TP_2}{FP_1 - FP_2} = \frac{p(-)C(+|-)}{p(+)C(-|+)} \quad (1)$$

where $p(a)$ is the probability of a given example being in class a, and $C(a|b)$ is the cost incurred if an example in class b is misclassified as being in class a. Equation 1 defines the gradient of an iso-performance line [9]. Classifiers sharing a line have the same expected cost for the ratio of priors and misclassification costs given by the gradient.

Even a single classifier can form an ROC curve. The solid line in figure 3 is produced by simply combining classifier B with the trivial classifiers: point (0,0) represents classifying all examples as negative; point (1,1) represents classifying all points as positive. The slopes of the lines connecting classifier B to (0,0) and to (1,1) define the range of the ratio of priors and misclassification costs for which classifier B is potentially useful, its operating range. For probability-cost ratios outside this range, classifier B will be outperformed by a trivial classifier. As with the single classifier, the operating range of any vertex on an ROC convex hull is defined by the slopes of the two line segments connected to it.

Thus the ROC representation allows an experimenter to see quickly if one classifier dominates another. Using the convex hull, potentially optimal classifiers and their operating ranges can be identified.

2.2 The Dual Representation

One of the questions posed in the introduction is how to determine the difference in performance of two ROC curves. For instance, in figure 3 the dashed curve is certainly better than the solid one. To measure how much better, one might be tempted to take the Euclidean distance normal to the lower curve. But this would be wrong on two counts. Firstly, the difference in expected cost is the weighted Manhattan distance between two classifiers, given by equation 2, not the Euclidean distance.

$$\begin{aligned} E[C_1] - E[C_2] &= (TP_1 - TP_2)\underbrace{p(+)C(-|+)}_{w_+} \\ &+ (FP_1 - FP_2)\underbrace{p(-)C(+|-)}_{w_-} \end{aligned} \quad (2)$$

Secondly, the performance difference should be measured between the appropriate classifiers on each ROC curve. When using the convex hull these are the best classifiers for the particular cost and class frequency defined by the weights w_+ and w_- in equation 2. In figure 3 for a probability-cost ratio of say 2.1 the classifier marked A on the dashed curve should be compared to the one marked B on the solid curve. But if the ratio was 2.3, it should be compared to the trivial classifier marked C on the dashed curve at the origin. This is the classifier that always labels instances negative.

To directly compare the performance of two classifiers we transform an ROC curve into a cost curve. Figure 4 shows the cost curves corresponding to the ROC curves in figure 3. The x-axis in a cost curve is the probability-cost function for positive examples, $PCF(+) = w_+/(w_+ + w_-)$ where w_+ and w_- are the weights in equation 2. This is simply $p(+)$, the probability of a positive example, when the costs are equal. The y-axis is expected cost normalised with respect to the cost incurred when every example is incorrectly classified. The dashed and solid cost curves in figure 4 correspond to the dashed and solid ROC curves in figure 3. The horizontal line atop the solid cost curve corresponds to the classifier marked B. The end points of the line indicate the classifier's operating range ($0.3 \leq PCF(+) \leq 0.7$), where it

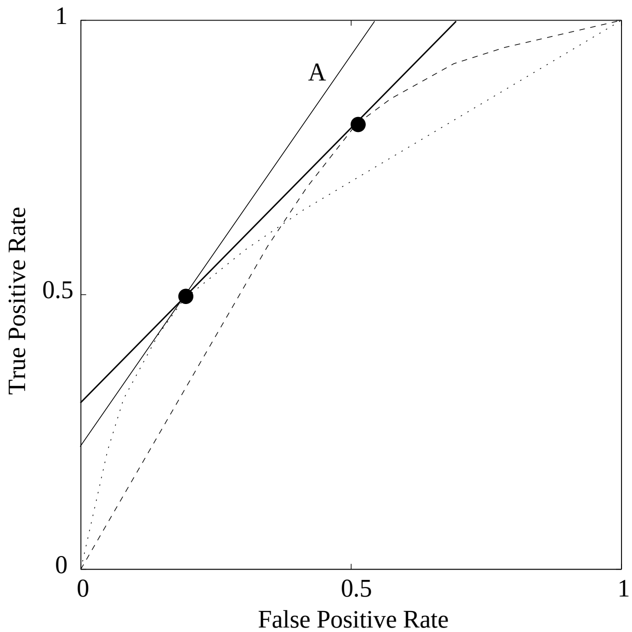

Figure 5: ROC Space Crossover

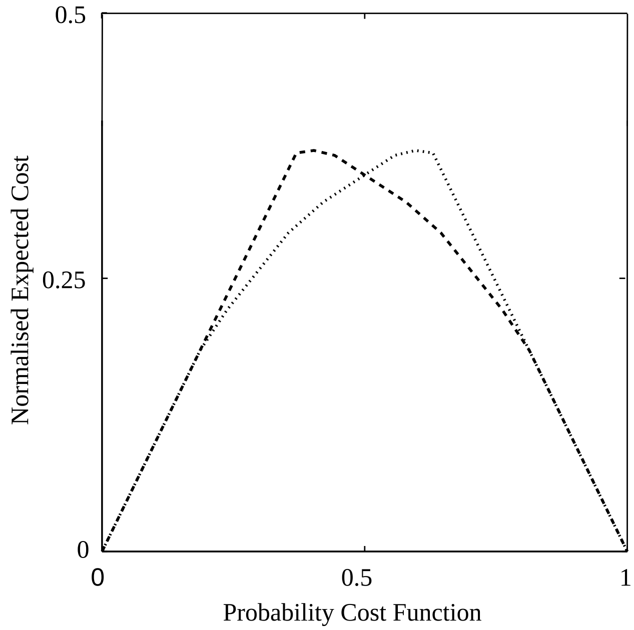

Figure 6: Cost Space Crossover

outperforms the trivial classifiers. It is horizontal because $FP = 1 - TP$ for this classifier (see below). At the limit of its operating range this classifier's cost curve joins the cost curve for the majority classifier. Each line segment in the dashed cost curve corresponds to one of the points (vertices) defining the dashed ROC curve.

The distance between cost curves for two classifiers directly indicates the performance difference between them. The dashed classifier outperforms the solid one – has a lower or equal expected cost – for all values of $PCF(+)$. The maximum difference is about 20% (0.25 compared to 0.3), which occurs when $PCF(+)$ is about 0.3 (or 0.7). Their performance difference is negligible when $PCF(+)$ is near 0.5, less than 0.2 or greater than 0.8.

It is certainly possible to get all this information from the ROC curves, but it is not trivial. The gradients of lines incident to a point must be determined to establish its operating range. To calculate the difference in expected cost, an iso-performance line must be brought into contact with each convex hull to determine which points must be compared. To find the actual costs the weighted Manhattan distance between them must be calculated. All this information is explicit in the alternative representation.

The second question posed in the introduction was for what range of cost and class distribution is one classifier better than another. Suppose we have the two hulls in ROC space, the dotted and dashed curves of figure 5. The solid lines indicate iso-performance lines. The line designated A touches the convex hull indicated by the dotted curve. A line with the same slope touching the other hull would be lower and to the right and therefore of higher expected cost. If we roll this line around the hulls until it touches both of them we find points on each hull of equal expected cost, for a particular cost or class frequency. Continuing to roll the line shows that the hull indicated by the dashed line becomes the better classifier. It is noteworthy that the crossover point of the two hulls says little about where one curve outperforms the other. It only denotes where both curves have a classification performance that is the same but suboptimal for any costs or class frequencies. Figure 6 shows the cost graph that is the dual of the ROC graph of figure 5. Here it can immediately be seen that the dotted line has a lower expected cost and therefore outperforms the dashed line to the left of the crossover point and vice versa. This crossover point when converted to ROC space becomes the line touching both hulls shown in figure 5.

2.2.1 Constructing the Dual Representation

To construct the alternative representation we use the normalised expected cost. The expected cost of a classifier is given by equation 3.

$$E[C] = (1 - TP)p(+)C(-|+) + FPp(-)C(+|-) \qquad (3)$$

The worst possible classifier is one that labels all instances incorrectly so $TP = 0$ and $FP = 1$ and its expected cost is given by equation 4.

$$E[C] = p(+)C(-|+) + p(-)C(+|-) \qquad (4)$$

The normalised expected cost is then produced by dividing the right hand side of equation 3 by that of equation 4 giving equation 5.

$$NE[C] = \frac{(1-TP)p(+)C(-|+) + FPp(-)C(+|-)}{p(+)C(-|+) + p(-)C(+|-)} \quad (5)$$

Then replacing the normalised probability-cost terms with the probability-cost function $PCF(a)$ as in equation 6 results in equation 7.

$$PCF(a) = \frac{p(a)C(a|\overline{a})}{p(+)C(-|+) + p(-)C(+|-)} \quad (6)$$

$$NE[C] = (1-TP) * PCF(+) + FP * PCF(-) \quad (7)$$

Because $PCF(+) + PCF(-) = 1$, we can rewrite equation 7 to produce equation 8 which is the straight line representing the classifier.

$$NE[C] = (1 - TP - FP) * PCF(+) + FP \quad (8)$$

A point (TP, FP) representing a classifier in ROC space is converted by equation 8 into a line in cost space. A line in ROC space is converted to a point in cost space, using equation 9, where S is the slope and TP_o the intersection with the true positive rate axis. Both these operations are invertible. So there is also a mapping from points (lines) in cost space to lines (points) in ROC space. Therefore there is a bidirectional point/line duality between the ROC and cost representations.

$$\begin{aligned} PCF(+) &= \frac{1}{1+S} \\ NE[C] &= (1-TP_o)PCF(+) \end{aligned} \quad (9)$$

Figure 7 shows lines representing four extreme classifiers in the cost space. At the top is the worst classifier, it is always wrong and has a constant normalised expected cost of 1. At the bottom is the best classifier, it is always right and has a constant cost of 0. The classifier that always chooses negative has zero cost when $PCF(+) = 0$ and a cost of 1 when $PCF(+) = 1$. The classifier that always chooses positive has cost of 1 when $PCF(+) = 0$ and a zero cost when $PCF(+) = 1$. Within this framework it is apparent that we should never use a classifier outside the shaded region of figure 7 as a lower expected cost can be achieved by using the majority classifier which chooses one or other of the trivial classifiers depending on $PCF(+)$.

At the limits of the normal range of the probability-cost function equation 8 simplifies to equation 10. To plot a classifier on the cost graph, we set the point on the left hand side y-axis to FP and the point on the right hand side y-axis to $(1-TP)$ and connect them by a straight line. Figure 8 shows a classifier with $FP = 0.09$ and $TP = 0.36$. The line represents the expected cost of the classifier over the full range of possible costs and class frequencies.

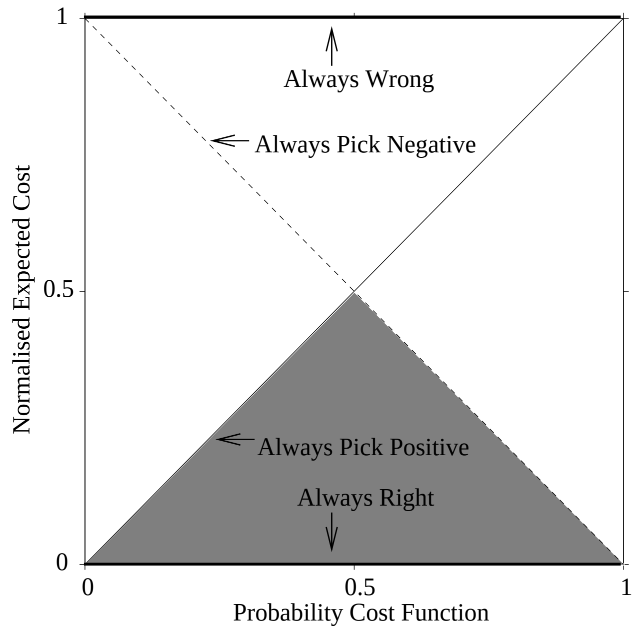

Figure 7: Extreme Classifiers

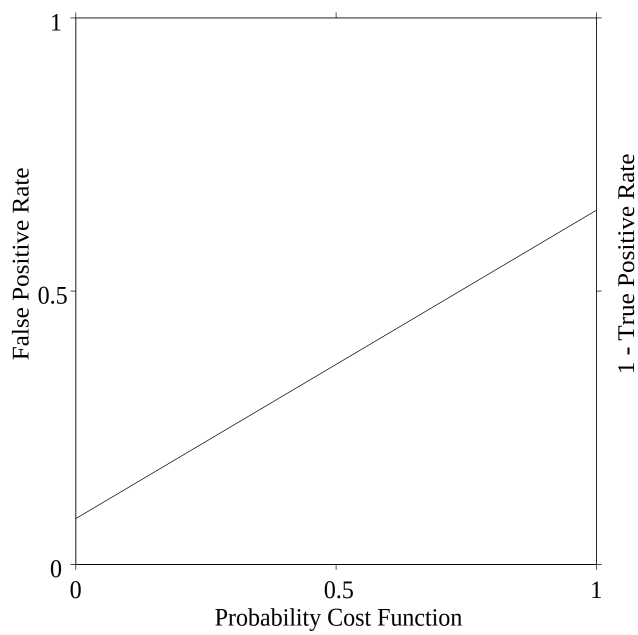

Figure 8: A Single Classifier

$$NE[C] = \begin{cases} FP, & \text{when } PCF(+) = 0 \\ (1-TP), & \text{when } PCF(+) = 1 \end{cases} \quad (10)$$

This procedure can be repeated for a set of classifiers, as shown in figure 9. We can now compare the difference in expected cost between any two classifiers. There is no need for the calculations required in the ROC space, we can directly measure the vertical height difference at some particular probability-cost value. Dominance is explicit in the

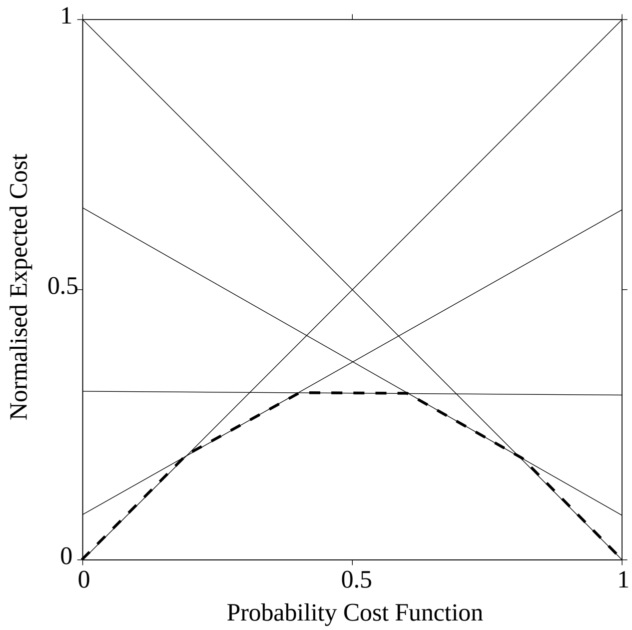

Figure 9: A Set of Classifiers

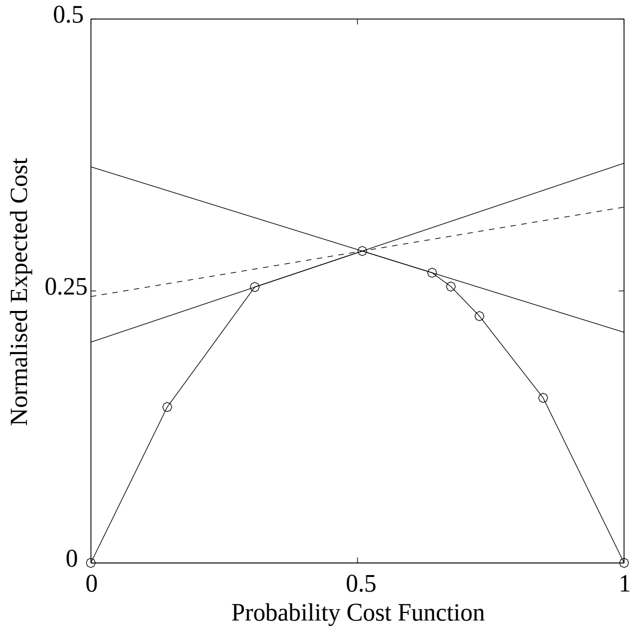

Figure 10: The Weighted Sum of Two Classifiers

cost space. If one classifier is lower in expected cost across the whole range of the probability-cost function, it dominates the other. Each classifier delimits a half-space. The intersection of the half-spaces of the set of classifiers gives the lower envelope indicated by the dashed line in figure 9. This effectively chooses the classifier that has the minimum cost for a particular operating range. This is equivalent to the upper convex hull in the ROC space. This equivalence arises from the duality of the two representations.

2.2.2 *Representing Other Performance Criteria*

In this section we look at how the other performance criteria discussed by Provost and Fawcett [10] are dealt with in cost space. They are as follows: error rate, area under the curve, Neyman-Pearson criterion and workforce utilisation.

As error rate is produced by setting all the costs in equation 5 to one, the cost graph is easily turned into an accuracy graph. The vertical distance between curves would then represent the difference in accuracy. There is no direct mapping of area under the curve in ROC space to cost space. But we can measure area under the curve in cost space and it has an intuitive meaning. Let us assume we do not know the probability-cost value used in practice, but we will use the appropriate classifier on the lower envelope when it is known. The area under the curve is the expected cost, assuming a uniform distribution $p(x)$ where x is the probability-cost value (the x-axis in the cost graph). Indeed if the probability distribution $p(x)$ is known the expected cost can be determined using equation 11. This also allows a comparison of two classifiers, or lower envelopes, where one does not strictly dominate the other. The difference in area under the two curves gives the expected advantage of using one classifier over another.

$$TEC = \int_o^1 NE[C(x)]p(x)dx \qquad (11)$$

A point on an edge of the ROC convex hull is not one of the original classifiers, but it can be realised by combining the two classifiers incident to it in a probabilistic way [10]. The probabilistic weighting is determined by the distance of the point to each classifier. As the cost graph is a dual representation to the ROC graph, there are also duals to operations, such as averaging two classifiers. In the cost graph, the combined classifier is a line, shown as the dotted line in figure 10. This is just the weighted sum of the two classifiers on the lower envelope, indicated by the solid lines, that intersect at a given vertex.

This becomes important when considering criteria such as Neyman-Pearson and workforce utilisation. The Neyman-Pearson criterion comes from statistical hypothesis testing and minimises the probability of a type two error for a maximum allowable probability of a type one error. For our purposes, this determines the maximum false positive rate and the aim is then to find the classifier with the largest true positive rate. This can be readily found on an ROC hull by drawing a vertical line for the particular value of FP, as shown by the dashed line in figure 11. The maximum value of TP (the minimum probability of a type two error) is where the line intersects the hull.

The procedure is very similar in the cost space. Remembering that the intersection of a classifier with the y-axis gives the false positive rate, then a point can be placed on the axis representing the criterion. This is marked FP in figure 12. Immediately on either side of this point are the equivalent points of two of the classifiers forming sides of the lower envelope. Connecting the new point to where the two

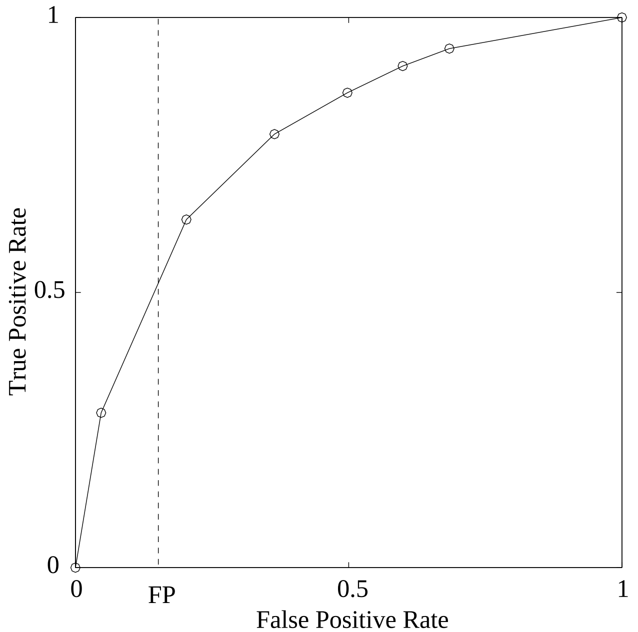

Figure 11: ROC Curve: Neyman-Pearson Criterion

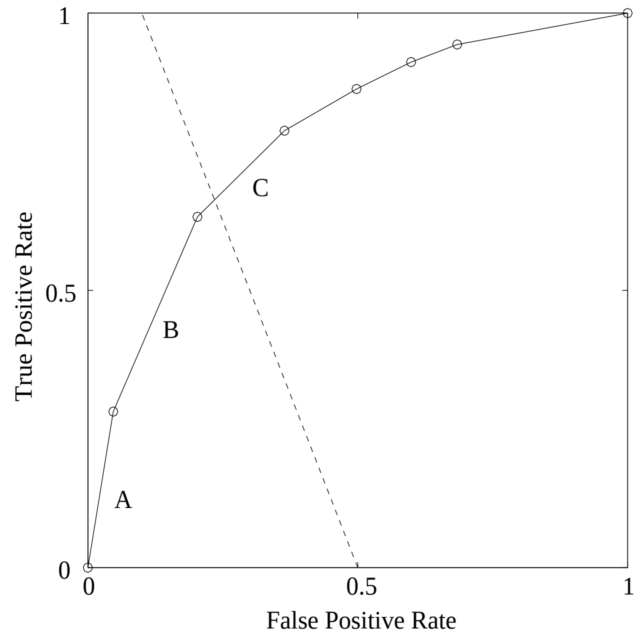

Figure 13: ROC Curve: Workforce Utilisation

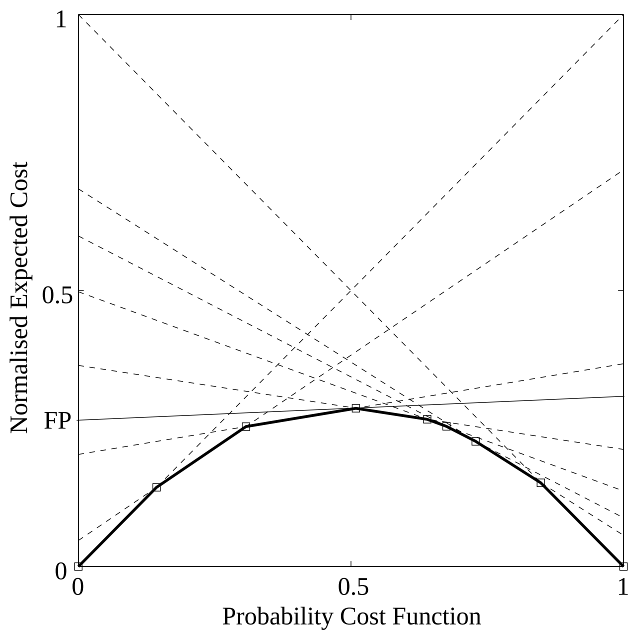

Figure 12: Cost Curve: Neyman-Pearson Criterion

classifiers intersect automatically gives the classifier meeting the Neyman-Pearson criterion.

Unfortunately, although the workforce utilisation criterion can be dealt with in cost space, it does not have the simple visual impact apparent in ROC space. The workforce utilisation criterion is based on the idea that a workforce can handle a fixed number of cases, factor C in equation 12. To keep the workforce maximally busy we want to select the best C cases, achieved by maximising the true positive rate. This is realised by the equality condition of equation 12 and is the line given by equation 13, such as the dashed line in figure 13. This line will be transformed to a point in the cost graph using equation 9 and is shown as the small circle on the left hand side of figure 14. The line's slope is negative, resulting in a $PCF(+)$ outside the normal interval of zero to one. We might consider it a virtual point, but strictly there is no constraint $PCF(a) \geq 0$ and so this represents a valid point on the line representing the classifier.

$$TP * P + FP * N \leq C \tag{12}$$

$$TP = -\frac{N}{P} * FP + \frac{C}{P} \tag{13}$$

The Neyman-Pearson criterion can be considered a special case of workforce utilisation, when the constraint only involves false positives. So for workforce utilisation a similar procedure to the one discussed above could be used for finding the appropriate classifier. All that would be required is to extend the original classifiers out until they have the same $PCF(+)$ value as the virtual point. Unfortunately this point may be arbitrarily far outside the normal range, which militates against easy visualisation. So instead below we give a simple algorithmic solution.

To solve the problem algorithmically in ROC space, a walk along the sides A, B, C of the convex hull, shown in figure 13, would be used to find the intersection point with the constraint. At each step, the edge is extended into a line and its intersection point with the constraint is tested to see if is between the two vertices, representing classifiers, that define the side. Equivalently in cost space we walk a line connected to the virtual point along the vertices A, B,

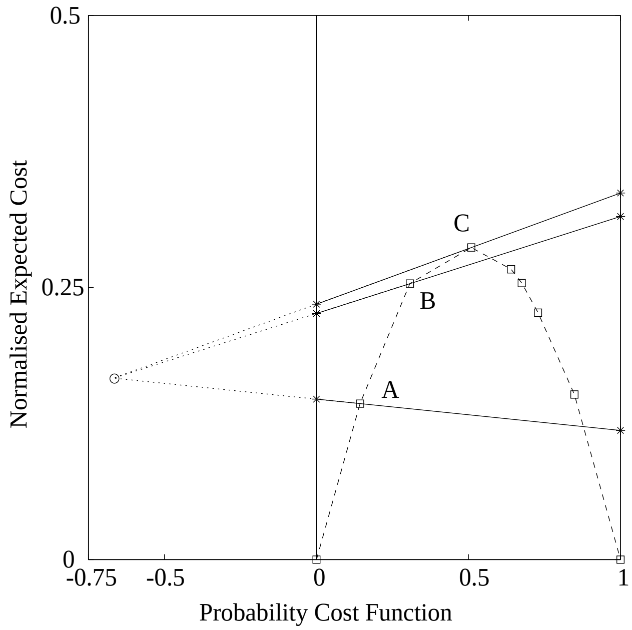

Figure 14: Cost Curve: Workforce Utilisation

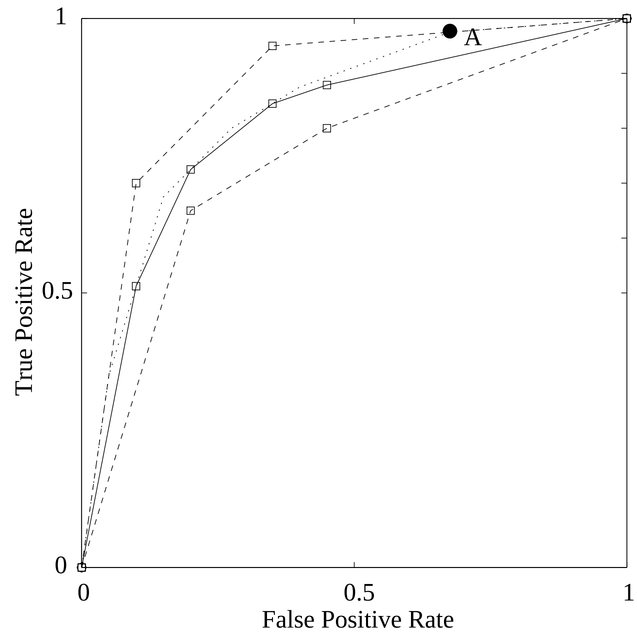

Figure 15: Average ROC Curves

C of the lower envelope, shown in figure 14. At each step, the slope of the line is tested to see if is between the two lines, representing classifiers, sharing the same vertex. In both spaces the appropriate classifier is found when the test is successful. In cost space, virtual points can be avoided if we rearrange the terms of equation 13 and substitute for the gradient of equation 8 resulting in equation 14. This can be solved for each point on the lower envelope. So a walk along vertices A, B, C of the lower envelope would produce the classifiers represented by the solid lines in figure 14, spanning just the normal probability-cost values.

$$NE[C] = \left(1 - \frac{C}{P} + (\frac{N}{P} - 1)FP\right) PCF(+) + FP \qquad (14)$$

In this section we have shown that the cost graph, can represent most of the alternative metrics discussed by Provost and Fawcett [10]. This is not surprising given the duality between the two spaces. But the different representations have different intuitive appeal. Certainly for the direct representation of costs, the cost graph seems the most intuitive. However we have also seen that for some metrics like the workforce utilisation criterion the ROC graph provides better visualisation.

2.2.3 Averaging Multiple Curves

Figure 15 shows two ROC curves, in fact convex hulls, represented by the dashed lines. If these are the result of training a classifier on different random samples, or some other cause of random fluctuation in the performance of a single classifier, their average can be used as an estimate of the classifier's expected performance. There is no universally agreed-upon method of averaging ROC curves. Swets and Pickett [13] suggest two methods, pooling and "averaging", and Provost et al. [11] propose an alternative averaging method.

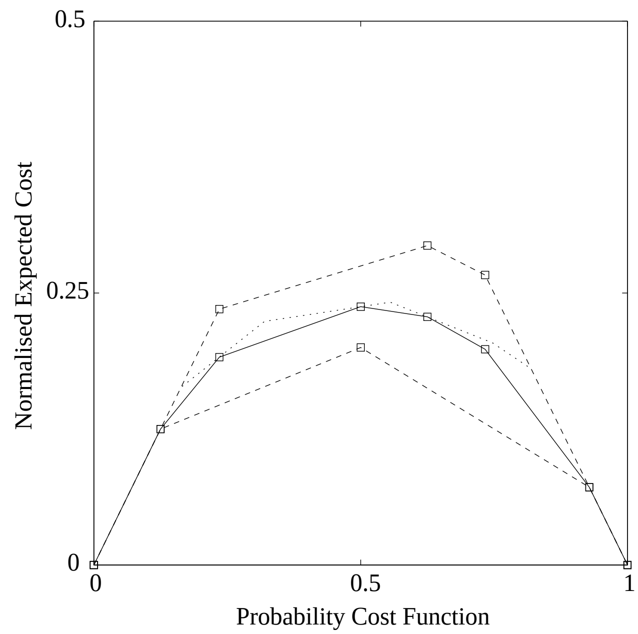

Figure 16: Average Cost Curves

The Provost et al. method is to regard y, here the true positive rate, as a function x, here the false positive rate, and to compute the average y value for each x value. This average is shown as a solid line in figure 15, with each vertex corresponding to a vertex from one or other of the dashed curves. Figure 16 shows the equivalent two cost curves, lower envelopes, represented by the dashed lines. The solid line is the result of the same averaging procedure but y and x are now the cost space axes. If the average curve in ROC space is transformed to cost space the dotted line results. Similarly, the dotted line in figure 15 is the result of transforming the average cost curve into ROC space. The curves are not the same.

The reason these averaging methods do not produce the same result is that they differ in how points on one curve are put into correspondence with points on the other curve. For the ROC curves points correspond if they have the same FP value. For the cost curves points correspond if they have the same $PCF(+)$ value, i.e. when $PCF(+)$ is in both their operating ranges. It is illuminating to look at the dotted line in the top right hand corner of figure 15. The vertex labelled "A" is the result of averaging a non-trivial classifier on the upper curve with a trivial classifier on the lower curve. This average takes into account the operating ranges of the classifiers and is significantly different from a simple average of the curves.

The cost graph average has a very clear meaning, it is the average normalised expected cost assuming that the classifier used for a given $PCF(+)$ value is the best available one. Notably the Provost et al. ROC averaging method, indicated by the dotted curve in figure 16, gives higher normalised expected costs for many $PCF(+)$ values. This is due to the average including at least some suboptimal classifiers. Pooling, or other methods of averaging ROC curves (e.g. choosing classifiers based on TP), will all produce different results, and all give higher normalised expected costs compared to the cost graph averaging method.

When estimating the expected performance of a classifier the average should be based on the selection procedure i.e. how the curve will ultimately be used to select an individual classifier. So far, we have compared curves without explicitly mentioning a selection procedure, but implicitly we are assuming the selection procedure inherent in using the lower envelope of the cost graph and the ROC convex hull: the point selected is the one that is optimal for the given $PCF(+)$ value. In this case the average based on the normalised expected cost is appropriate. This does not mean however that other averages are incorrect. Each is based on a different selection procedure which will be appropriate for different performance criteria. Provost et al.'s averaging method is appropriate when the performance criterion calls for classifier selection based on FP, such as the Neyman-Pearson criterion.

2.3 A Suboptimal Selection Procedure

We have just seen that different averages of two curves result from different selection procedures, due to the different ways of deciding which point on one curve will correspond to which point on another curve. A selection procedure is also necessary to compare two curves quantitatively, since by its very nature quantitative comparison involves summing the difference in performance of corresponding points.

The selection method that is most commonly used in comparing learning algorithms is parameter-based. For example, suppose one wishes to compare two learning algorithms and that an ROC curve is generated for each algorithm by undersampling or oversampling to create various class ratios in the training set. Typically, one would compare the performance

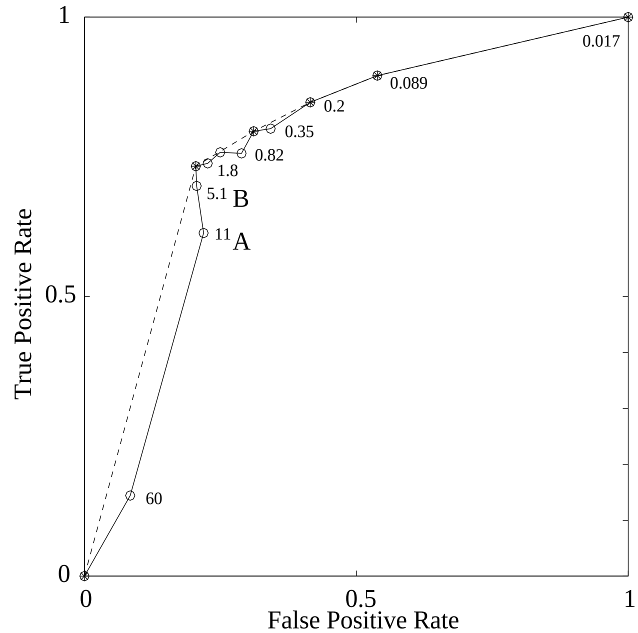

Figure 17: ROC for Sonar Data

of the classifiers produced on the same training sets: this is choosing which points on each curve correspond based on the underlying parameter that generated them rather than on their operating range. It might happen that algorithm A trained with a 5:1 ratio produces a classifier with the same operating range as the classifier produced by algorithm B with a 10:1 ratio. This could only be determined by looking at the convex hull or the lower envelope in their respective spaces.

The fact that the optimal classifier for a particular $PCF(+)$ value is not necessarily the one produced by a training set with the same $PCF(+)$ characteristics is illustrated in figure 17, which shows ROC curves for the sonar data set from the UCI collection [1]. The points represented by circles, and connected by solid lines, were generated using C4.5 (release 7 using information gain) modified to account for costs (by altering the values inside C4.5 representing priors). Each point is marked with the probability-cost ratio used to produce it. If the probability-cost ratio is 11 at the time of application, for example, parameter-based selection would select classifier A, since it was produced by a training set with a 11:1 ratio.

Using the convex hull selection method, the dashed line in figure 17, classifiers would be selected according to the slope of its sides. This would result in the expected cost shown by the lower envelope, the dashed line in figure 18. If instead, the classifiers are chosen according to the probability-cost ratio input to the classifier, the solid line is produced. A probability-cost ratio R is converted to a $PCF(+)$ value using the $PCF(+) = 1/(1 + R)$. In cost space, classifier A will be chosen when $PCF(+) = 1/(1 + 11)$ and classifier B when $PCF(+) = 1/(1 + 5.1)$, as shown in figure 18. Changing from classifier A to classifier B we assume occurs at the mid-point of these two probability-cost values. The area between this curve and the lower envelope is a measure

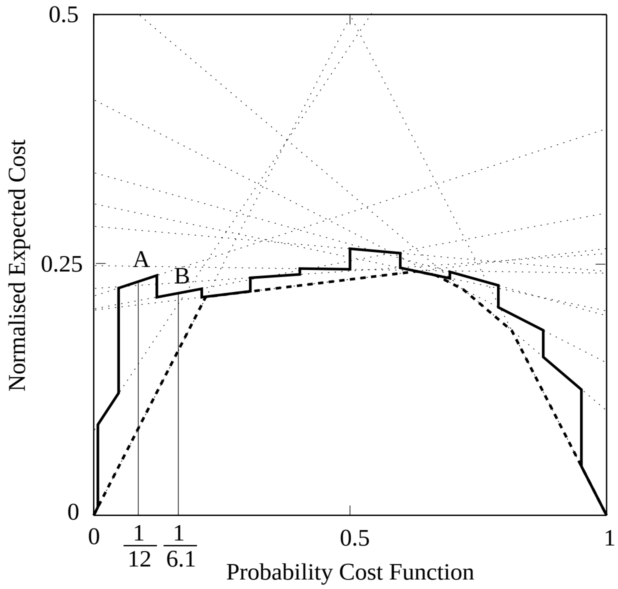

Figure 18: Cost for Sonar Data

of the additional cost of using this selection procedure over the optimal one. The large difference at the left hand and right hand sides is due to not using the majority classifier at the appropriate time. This shows the clear disadvantage of using a classifier outside its operating range.

3. LIMITATIONS AND FUTURE WORK

One limitation of this work, which is common to that of ROC analysis, is that we have not investigated the situation of more than two classes. Although the ideas should readily extend to three or more classes, the main advantage of this approach is it ease of human understandability. Higher dimensional functions are notoriously difficult to visualise and the number of dimensions increases quadratically with the number of classes. Due to the duality between the two representations there might be little merit in using one over the other in this situation. However, if the high dimensional space can be projected into a two dimensional space, the improved understandability would again be an advantage. Another limitation is that we have not investigated other commonly used metrics for evaluating classifier performance such as lift. One interesting avenue of future research is whether or not there are alternative dualities based on such metrics.

4. CONCLUSIONS

This paper has demonstrated an alternative to ROC analysis, which represents the cost explicitly. It has shown there is a point/line duality between the two representations. This allows the cost representation to maintain many of the ROC representation's advantages, while making notions such as operating range visually clearer. It also allows the easy calculation of the quantitative difference between classifiers. The fact that the two representations are dual representations makes it unnecessary to choose one over the other, as we have shown it is easy to switch between the two.

5. ACKNOWLEDGEMENTS

We would like to thank the reviewers for their valuable suggestions and the Natural Sciences and Engineering Research Council of Canada for financial support.

6. REFERENCES

[1] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, University of California, Irvine, CA . www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, CA, 1984.

[3] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 155–164, Menlo Park, CA, 1999. AAAI Press.

[4] T. Fawcett and F. Provost. Combining data mining and machine learning for effective user profiling. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 8–13, Menlo Park, CA, 1996. AAAI Press.

[5] T. Fawcett and F. Provost. Adaptive fraud detection. *Journal of Data Mining and Knowledge Discovery*, 1:195–215, 1997.

[6] M. Kubat, R. C. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30:195–215, 1998.

[7] L. B. Lusted. *Introduction to Medical Decision Making.* Charles C. Thomas, Springfiled, Illinois, 1968.

[8] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 217–225, San Francisco, 1997. Morgan Kaufmann.

[9] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48, Menlo Park, CA, 1997. AAAI Press.

[10] F. Provost and T. Fawcett. Robust classification systems for imprecise environments. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 706–713, Menlo Park, CA, 1998. AAAI Press.

[11] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 43–48, San Francisco, 1998. Morgan Kaufmann.

[12] J. A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988.

[13] J. A. Swets and R. M. Pickett. *Evaluation of diagnostic systems : methods from signal detection theory.* Academic Press, New York, 1982.