Explicitly Representing Expected Cost: An Alternative to ROC Representation

Chris Drummond and Robert C. Holte

School of Information Technology and Engineering

University of Ottawa, 150 Louis Pasteur, P.O. Box 450 Stn A

Ottawa, Ontario, Canada, K1N 6N5

cdrummon,holte@site.uottawa.ca

Contact Author: Chris Drummond

Telephone Number: (613) 233-5161 Canada

Paper ID: 331

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 techniques used in ROC analysis to be readily reproduced in the cost space.

1 Introduction

Provost and Fawcett (1997) 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 in a dataset is to raise the cost of misclassifying the minority class. 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 introduced the idea of using a convex hull to find the best classifier for a particular combination of costs and class frequencies.

The ROC representation makes the trade-off between types of error readily apparent and is particularly useful when the costs or class frequencies are not known until the time of application. In Fawcett and Provost's (1996, 1997) 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 (Kubat, Holte, & Matwin, 1998) 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

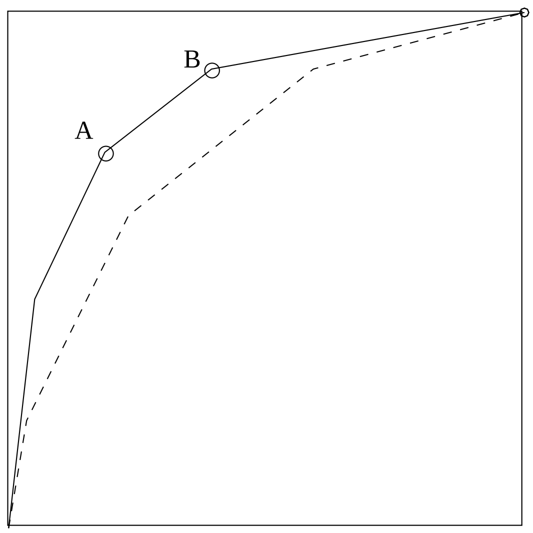

Figure 1: Comparing Performance

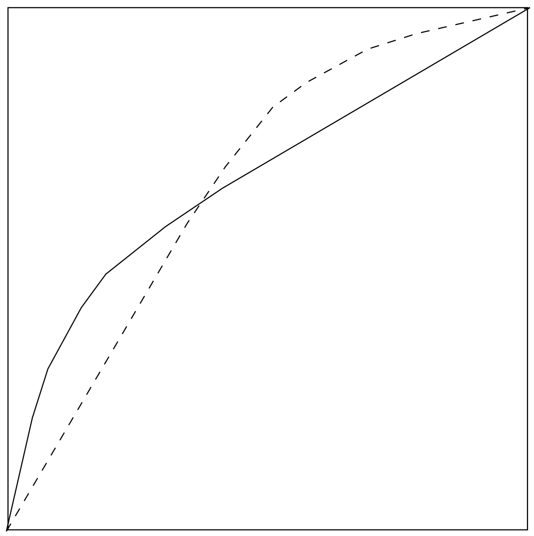

Figure 2: Performance Ranges

reports. For this purpose despite all of the strengths of the ROC representation, we found the graphs produced were not always easy to interpret. For instance, although it is easy to see which curve is better in figure 1, it is much harder to determine by how much. It 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 necessarily 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.

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 (1997) 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 (1988) 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 (Domingos, 1999; Pazzani, Merz, Murphy, Ali, Hume, & Brunk, 1997). 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 (Breiman, Friedman, Olshen, & Stone, 1984). 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 (Domingos, 1999).

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 (1988) 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 (Provost & Fawcett, 1997). It has further been shown that dominance implies superior performance for a variety of commonly-used performance measures (Provost & Fawcett, 1998). 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.

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

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 (Provost & Fawcett, 1997). 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.

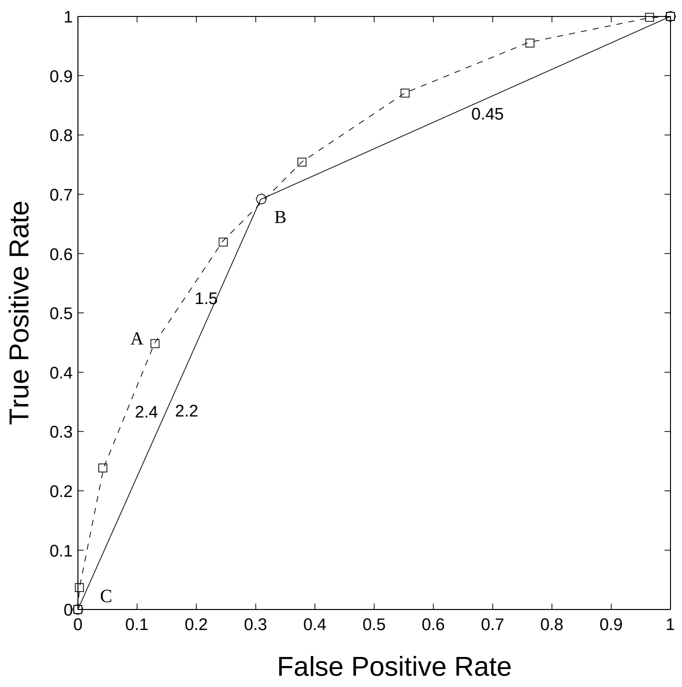

Figure 3: Comparing Two ROC curves

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.

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

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. 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, $PCF(+) = w_+/(w_+ + w_-)$ which is simply $p(+)$ when the costs are one. 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 region ($0.3 \leq PCF(+) \leq 0.7$). 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 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.

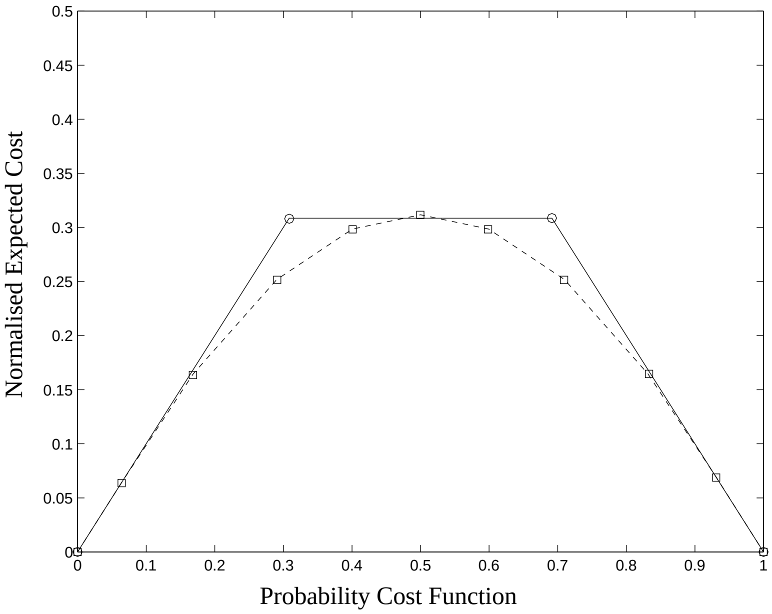

Figure 4: Comparing Misclassification Costs

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.

2.2.1 Constructing the Dual Representation

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

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

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.

$$E[C] = p(+)C(-|+) + p(-)C(+|-) \tag{4}$$

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

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

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

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

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

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

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

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.

$$PCF(+) = \frac{1}{1 + S} \quad and \quad NE[C] = (1 - TP_o)PCF(+) \tag{9}$$

Figure 5 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

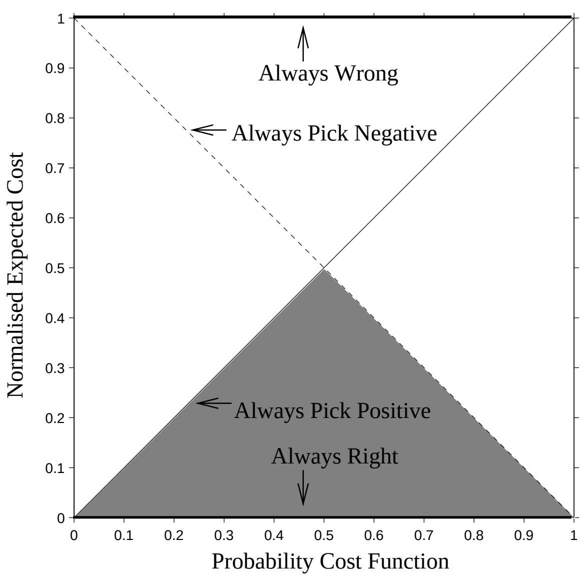

Figure 5: Extreme Classifiers

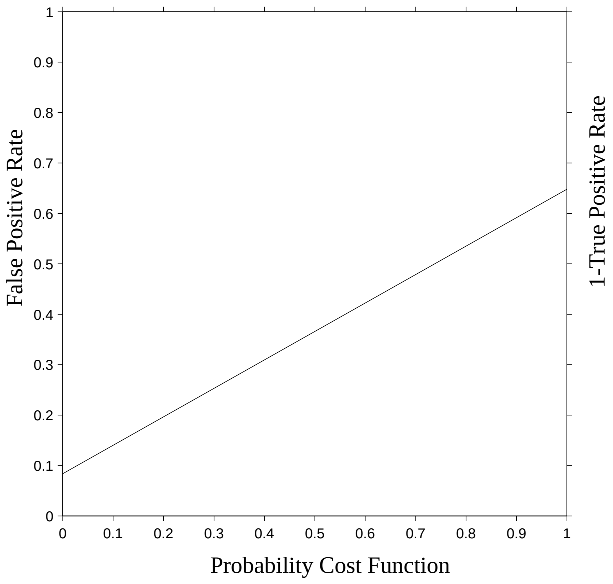

Figure 6: A Single Classifier

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 5 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 6 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.

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

This procedure can be repeated for a set of classifiers, as shown in figure 7. 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 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 7. 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.

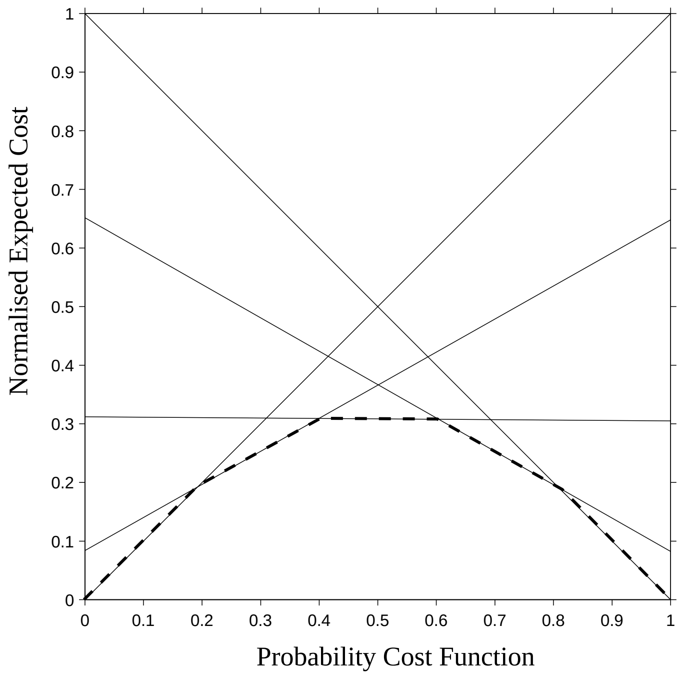

Figure 7: A Set of Classifiers

2.2.2 Representing Other Performance Criteria

In this section we look at how the other performance criteria discussed by Provost and Fawcett (1998) 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 represents 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.

$$TEC = \int_{o}^{1} NE[C(x)]p(x)dx \tag{11}$$

A point on a side 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 (Provost & Fawcett, 1998). 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. In the cost graph, the combined classifier is a line, shown as the dotted line in figure 8. 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 can be considered a special case of workforce utilisation, shown in equation 12, when the constraint only involves false positives. The best classifier under this criterion is realised by the equality condition of equation 12, as this maximises the true positive rate. This is the line in given by equation 13, such as the dashed line in figure 9. 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 10. The line's slope is negative, resulting in a $PCF(+)$ outside the normal interval of zero to one. The resultant

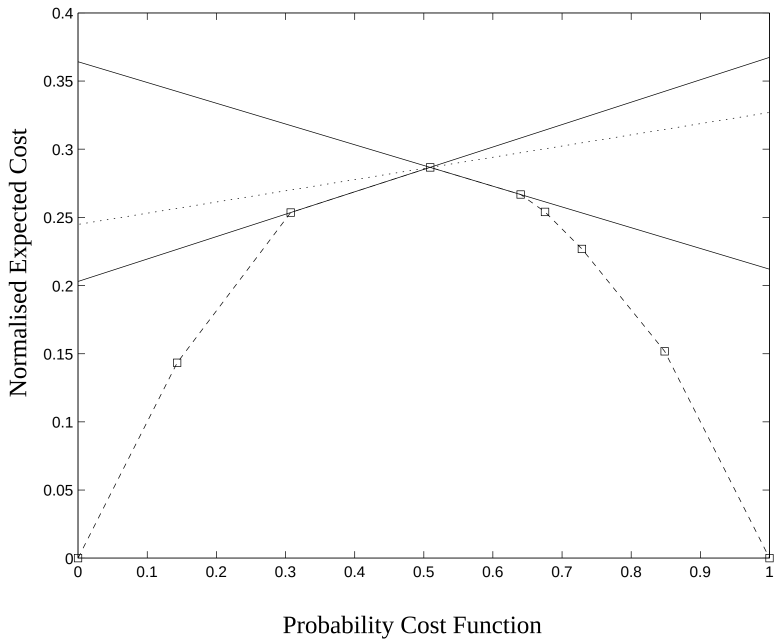

Figure 8: The Weighted Sum of Two Classifiers

point might be considered a virtual point as it exists outside the normal range of the graph. But strictly there is no constraint $PCF(a) \geq 0$ and so this represents a valid point on the line representing the classifier.

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

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

In ROC space, a walk along the sides A, B, C of the convex hull, shown in figure 9, would be used to find the intersection point with the constraint. At each step, the side 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, C of the lower envelope, shown in figure 10. 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 10, spanning just the normal probability-cost values.

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

In this section we have shown that the cost graph, can represent all the alternative metrics discussed by Provost and Fawcett (1998). 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. For other measures it may be the ROC graph.

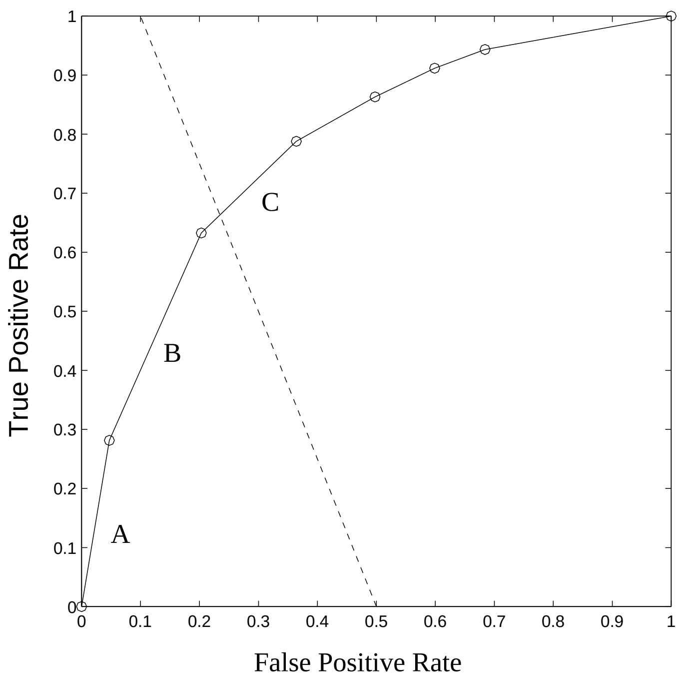

Figure 9: ROC Curve with Workforce Utilisation Constraint

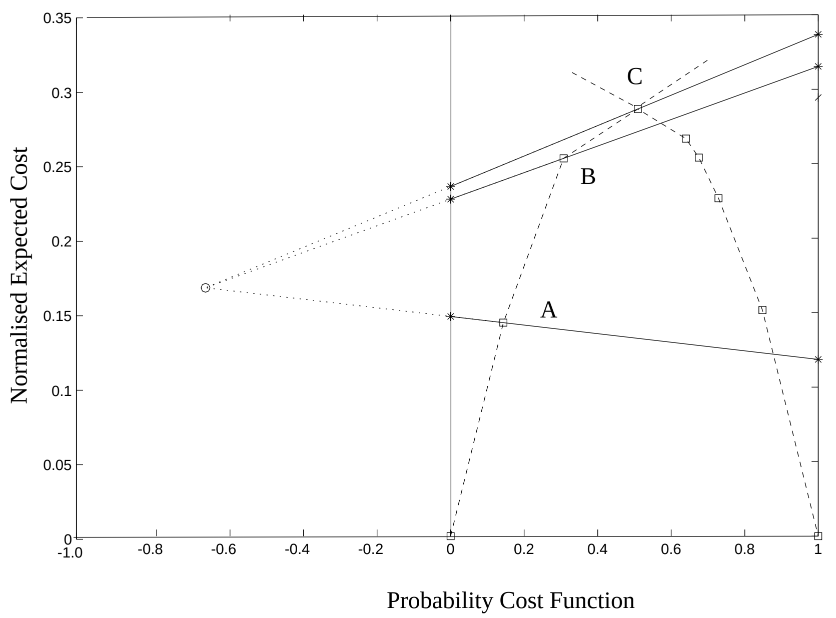

Figure 10: Cost Curve with Workforce Utilisation Constraint

2.2.3 Performance Under Different Selection Procedures

In section 2.1 we discussed the idea of one curve dominating another in ROC space. If a curve is dominated, we know that for every classifier on that curve there is a better one on the dominating curve. But if there is no guarantee that these two classifiers would be selected under the same criterion in practice, lower misclassification costs might not result. Thus implicit in the notion of dominance is a procedure for selecting a classifier from each curve. It is not strictly necessary in ROC analysis to fully define the selection procedure. It is only necessary to establish a point to point correspondence between the curves. To produce a meaningful curve in cost space, we do require a completely defined selection procedure. This is necessary in either space if any sort of quantitative comparison of the performance of different classifiers is to be made. It is also necessary if the range where one classifier is better than another is to be established. A selection procedure is even needed if we want to say anything quantitative about the performance of a single classifier.

What sorts of selection procedure might be used? The convex hull is one such procedure which chooses only the potentially optimal classifiers based on iso-performance lines. This is equivalent to using classifiers on the lower envelope in cost space. An alternative selection procedure would be to choose a classifier based on the misclassification costs given to a classifier. Figure 11 shows ROC curves for the sonar data set from the UCI collection (Blake & Merz, 1998). 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 cost ratio used to produce it. If the probability-cost ratio is 11 at the time application, for example, classifier A would be selected.

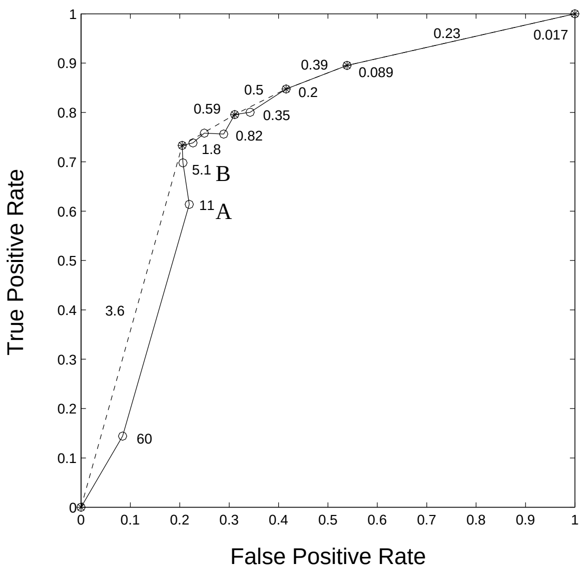

Figure 11: ROC for Sonar Data

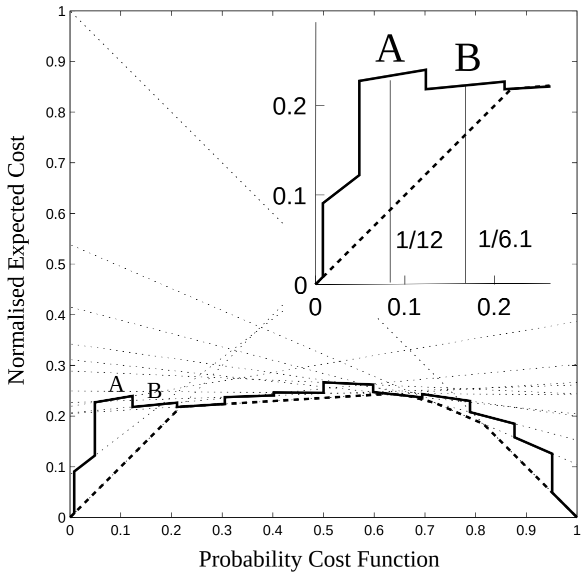

Figure 12: Cost for Sonar Data

Using the convex hull, the dashed line in figure 11, 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 12. If instead, the classifiers are chosen according to the misclassification costs input to the classifier, the solid line is produced. A probability cost ratio is converted to a probability cost 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 the insert of figure 12. 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 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.

A selection procedure is also important when determining the range where one classifier is better than another. Suppose we have the two hulls in ROC space, the dotted and dashed curves of figure 13. The solid lines indicate iso-perfomance 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. 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.

Figure 14 shows the cost graph that is the dual of the ROC graph of figure 13. 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 vica versa. This crossover point when converted to a line in ROC space, is the line touching both hulls shown in figure 13. Determining the crossover point in the cost space is not dependent on the curves being lower envelopes. It will be similar for any selection procedure based on probability-cost values.

A selection procedure must also be defined before we can average multiple curves, different procedures will produce different averages. One way to construct an average is to pool points from a single classifier over different samples. Then find the mean of each pool and construct the curve. This is appropriate if the selection procedure is based on some input parameter. Provost et al. (1998) argue that this averages points on different parts of an ROC curve. Instead they treat the curve simply as a function $TP = R(FP)$ and average the resulting functions. This is appropriate if the selection procedure is based on the false positive rate,

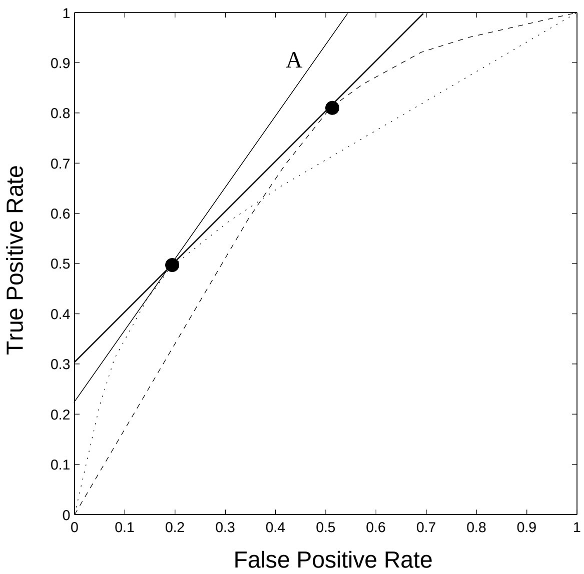

Figure 13: ROC Space Crossover

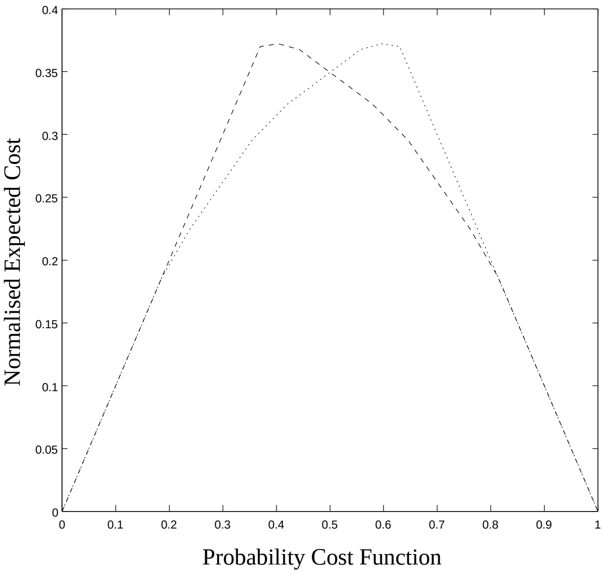

Figure 14: Cost Space Crossover

such as would be the case using the Neyman-Pearson criterion. A third approach is to select classifiers according to their operating ranges, but in ROC space it is unclear how to average the hulls. In cost space it is simply a matter of averaging the lower envelopes.

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.

References

Blake, C., & Merz, C. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]..

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth.

Domingos, P. (1999). MetaCost: A General Method for Making Classifiers Cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pp. 155–164.

Fawcett, T., & Provost, F. (1996). Combining Data Mining and Machine Learning for Effective User Profiling. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 8–13.

Fawcett, T., & Provost, F. (1997). Adaptive Fraud Detection. *Journal of Data Mining and Knowledge Discovery*, *1*, 195–215.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*, *30*, 195–215.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1997). Reducing Misclassification Costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 217–225.

Provost, F., & Fawcett, T. (1997). 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*, pp. 43–48.

Provost, F., & Fawcett, T. (1998). Robust Classification Systems for Imprecise Environments. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 706–713.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The Case Against Accuracy Estimation for Comparing Induction Algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 43–48.

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, *240*, 1285–1293.