What ROC Curves Can't Do (and Cost Curves Can)

Chris Drummond CHRIS.DRUMMOND@NRC.CA
Institute for Information Technology,
National Research Council Canada,
Ontario, Canada, K1A 0R6

Robert C. Holte HOLTE@CS.UALBERTA.CA
Department of Computing Science,
University of Alberta,
Edmonton, Alberta, Canada, T6G 2E8

Abstract

This paper shows that ROC curves, as a method of visualizing classifier performance, are inadequate for the needs of Artificial Intelligence researchers in several significant respects, and demonstrates that a different way of visualizing performance – the cost curves introduced by Drummond and Holte at KDD'2000 – overcomes these deficiencies.

1. Introduction

In this paper, our focus is on the visualization of a classifier's performance. This is one of the attractive features of ROC analysis – the tradeoff between false positive rate and true positive rate can been seen directly. A good visualization of classifier performance would allow an experimenter to immediately see how well a classifier performs and to compare two classifiers – to see when, and by how much, one classifier outperforms others.

We restrict the discussion to classification problems in which there are only two classes. The main point of this paper is to show that, even in this restricted case, ROC curves are not a good visualization of classifier performance. In particular, they do not allow any of the following important experimental questions to be answered visually:

- what is classifier C's performance (expected cost) given specific misclassification costs and class probabilities ?
- for what misclassification costs and class probabilities does classifier C outperform the trivial classifiers ?
- for what misclassification costs and class probabilities does classifier C1 outperform classifier C2 ?
- what is the difference in performance between classifier C1 and classifier C2 ?
- what is the average of performance results from several independent evaluations of classifier C (e.g. the results of 5-fold cross-validation) ?
- what is the 90% confidence interval for classifier C's performance ?

- what is the significance (if any) of the difference between the performance of classifier C1 and the performance of classifier C2 ?

The paper is organized around these questions. After a brief review of essential background material, there is a section devoted to each of these questions.

2. Background

For 2-class classification problems ROC space is a 2-dimensional plot with true positive rate (TP) on the y-axis and false positive rate (FP) on the x-axis. A single confusion matrix thus produces a single point in ROC space. An ROC curve is formed from a sequence of such points, including (0,0) and (1,1), connected by line segments. The method used to generate the sequence of points for a given classifier (or learning algorithm) depends on the classifier. For example, with Naive Bayes (Duda & Hart, 1973; Pazzani, Merz, Murphy, Ali, Hume, & Brunk, 1994) an ROC curve is produced by varying its threshold parameter. In the absence of any method to generate a sequence of ROC points a single classifier can form the basis of an ROC curve by connecting its ROC point to points (0,0) and (1,1).

An ROC curve implicitly conveys information about performance across all possible combinations of misclassifications costs and class distributions [1]. We use the term "operating point" to refer to a specific combination of misclassifications costs and class distributions.

One point in ROC space dominates another if it has a higher true positive rate and a lower false positive rate. If point A dominates point B, A will have a lower expected cost than B for all operating points. 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.

Cost curves were introduced in (Drummond & Holte, 2000). Performance (expected cost normalized to be between 0 and 1) is plotted on the y-axis. Operating points are plotted on the x-axis after being normalized to be between 0 and 1 by combining the parameters defining an operating point in the following way:

$$PCF(+) = \frac{p(+)\mathtt{C}(-|+)}{p(+)\mathtt{C}(-|+) + p(-)\mathtt{C}(+|-)} \tag{1}$$

where $\mathtt{C(-|+)}$is the cost of misclassifiying a positive example as negative, $\mathtt{C(+|-)}$is the cost of misclassifiying a negative example as positive, $p(+)$ is the probability of a positive example, and $p(-) = 1 - p(+)$. The motivation for this PCF definition, and cost curves more generally, originates in the simple situation when misclassification costs are equal. In this case $PCF(+) = p(+)$ and the y-axis becomes error rate, so the cost curve plots how error rate varies a function of the prevalance of positive examples. The PCF definition generalizes this idea to the case when when misclassification costs are not equal. The PCF formula is intimately tied to the definition of the slope of a line in ROC space, which plays a key role in ROC analysis. The x-axis of cost space is "slope in ROC space" normalized to be between 0 and 1 instead of being between 0 and infinity (historically this is how cost curves were invented).

1. "All" distributions and costs with certain standard restrictions. For class distributions "all" means any prior probabilities for the classes while keeping the class-conditional probabilities, or likelihoods, constant (Webb & Ting, 2004). For costs "all" means all combinations of costs such that a misclassification is more costly than a correct one.

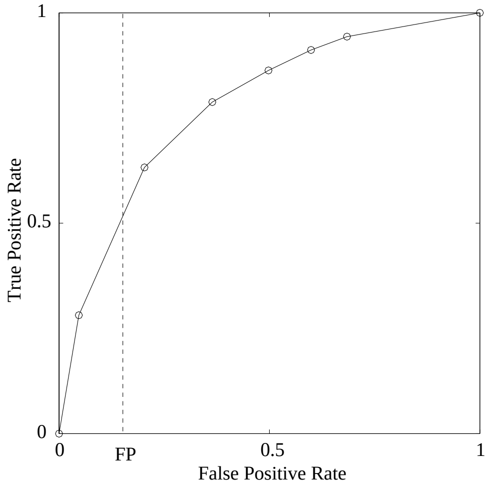

Figure 1: ROC curve

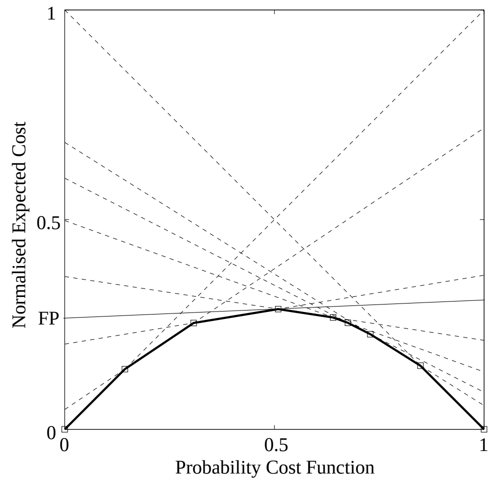

Figure 2: Corresponding Cost Curve

There is a point/line duality between ROC space and cost space, meaning that a point in ROC space is represented by a line in cost space, a line in ROC space is represented by a point in cost space, and vice versa. A classifier represented by the point (FP,TP) in ROC space is a line in cost space that has $y = FP$ when $x = 0$ and $y = 1 - TP$ when $x = 1$. The set of points defining an ROC curve become a set of lines in cost space For example, the ROC curve in Figure 1 consists of eight points (including (0,0) and (1,1)). Each point becomes a line in cost space, the eight dotted lines in Figure 2 (in both figures ignore the line labelled FP). Corresponding to the convex hull the points in ROC space is the lower envelope of the lines in cost space, indicated by the solid line in Figure 2.

3. Visualizing Classifier Performance

ROC analysis does not directly commit to any particular measure of performance. This is sometimes considered an advantageous feature of ROC curves. For example, Van Rijsbergen (1979) quotes Swets (1967) who argues that this is useful as it measures "discrimination power independent of any 'acceptable criterion' employed". Provost and Fawcett substantiate this argument by showing that ROC dominance implies superior performance for a variety of commonly-used performance measures (Provost & Fawcett, 1998). The ROC representation allows an experimenter to see quickly if one classifier dominates another and therefore, using the convex hull, to identify potentially optimal classifiers visually without committing to a specific performance measure.

Being independent of any particular performance measure can be a disadvantage when one has a particular performance measure in mind. ROC curves do not visually depict the quantitative performance of a classifier or the difference in performance between two classifiers. For example, Figure 3 shows the ROC curve for Naive Bayes using Gaussian probability estimators, on the sonar data set from the UCI collection. We can see immediately that the ROC curve is well above the chance line, the diagonal joining (0,0) to (1,1). One might conclude that this classifier's performance is "good", but there is no indication

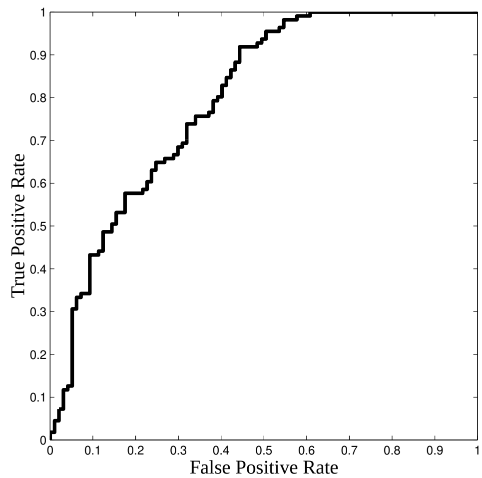

Figure 3: ROC Curve

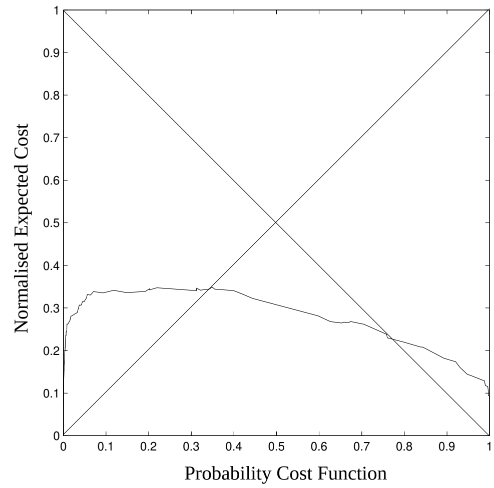

Figure 4: Cost Curve

as to how good it is. For example, we cannot tell from visual inspection what its error rate would be if misclassifications costs were equal and the two classes were equally likely.

By contrast, cost curves were defined to allow performance to be read off directly for any given operating point. The cost curve corresponding to the ROC curve in Figure 3 is the humped, mainly downsloping curve in Figure 4. We can directly read off the performance for any specific operating point and we can see how performance varies across the full range of possible operating points. For example, performance when misclassification costs are equal and the two classes are equally likely can be read off the plot by looking at the cost curve's value at $x = 0.5$. It is roughly 0.3, an error rate which is adequate but not especially "good". We can also see that performance does not vary much across the range of operating points: it is between 0.2 and 0.35 except when $PCF(+) > 0.9$.

4. Comparing a Classifier to the Trivial Classifiers

In an ROC diagram points (0,0) and (1,1) represent the trivial classifiers: (0,0) represents classifying all examples as negative, and (1,1) represents classifying all points as positive. The cost curves for these classifiers are the diagonal lines shown in Figure 4. The diagonal line from (0,0) to (1,1) is the cost curve for the classifier that classifies all examples as negative, and the diagonal line from (0,1) to (1,0) is the cost curve for the classifier that classifies all examples as positive.

The operating range of a classifier is the set of operating points where it outperforms the trivial classifiers. A classifier should not be used outside its operating range, since one can obtain superior performance by assigning all examples to a single class.

The operating range of a classifier cannot be seen readily in an ROC curve. It is defined by the slopes of the lines tangent to the ROC curve and passing through (0,0) and (1,1). By contrast, a classifier's operating range can be immediately read off of a cost curve: it is defined by the PCF values where the cost curve intersects the diagonal lines representing the trivial classifiers. For example, in Figure 4 it can be seen immediately that Naive Bayes performs worse than a trivial classifier when $PCF < 0.35$ or $PCF > 0.75$.

5. Choosing Between Classifiers

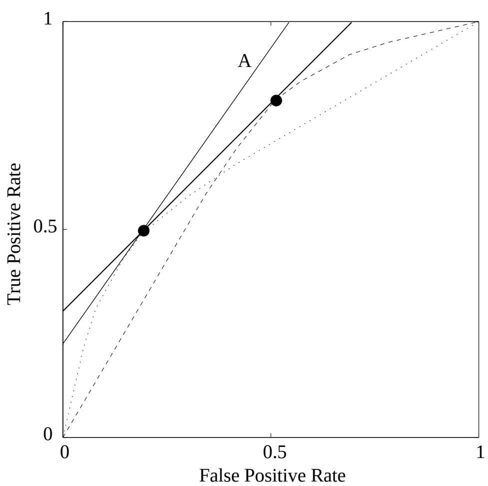

Figure 5: ROC Space Crossover

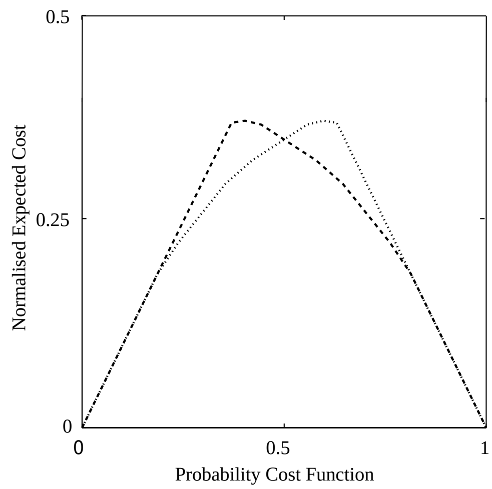

Figure 6: Cost Space Crossover

If the ROC curves for two classifiers cross, each classifier is better than the other for a certain range of operating points. Identifying this range visually is not easy in an ROC diagram and perhaps surprisingly the crossover point of the ROC curves has little to do with the range. Consider the ROC curves for two classifiers, the dotted and dashed curves of Figure 5. The solid lines indicate iso-performance lines. The line designated A touches the dotted ROC curve. A line with the same slope touching the other curve would be lower and to the right and therefore of higher expected cost. If we roll this line around the curves until it touches both of them we find points on each curve of equal expected cost. Continuing to roll the line shows that the dashed ROC curve becomes the better classifier.

Figure 6 shows the cost curves corresponding to the ROC curves in Figure 5. It can immediately be seen that the dotted line has a lower expected cost and therefore outperforms the dashed line when $PCF < 0.5$ and vice versa.

6. Comparing Classifier Performance

Figures 7 and 8 illustrate how much more difficult it is to compare classifiers with ROC curves than with cost curves. Although it is obvious in the ROC diagram that the dashed curve is better than the solid one, it is not easy, visually, to determine by how much. One might be tempted to take the Euclidean distance normal to the lower curve to measure the difference. But this would be wrong on two counts. Firstly, the difference in expected cost is the weighted Manhattan distance between two classifiers not the Euclidean distance.

Secondly, the performance difference should be measured between the appropriate classifiers on each ROC curve – the classifiers out of thet set of possibilities on each curve that would be used at each given operating point. To illustrate how intricate this is, suppose

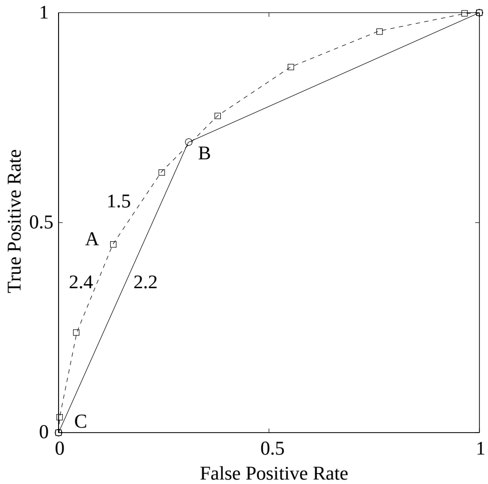

Figure 7: Comparing ROC Curves

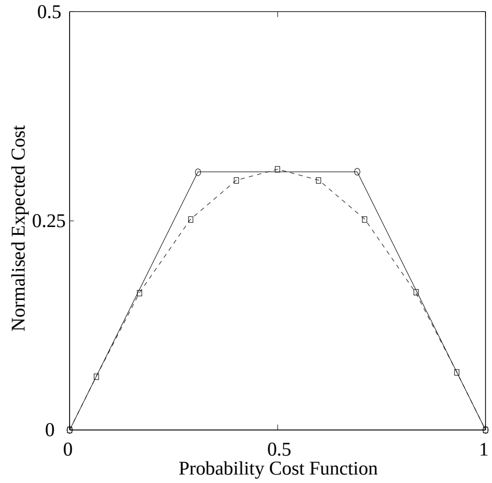

Figure 8: Comparing Cost Curves

the two classes are equiprobable but that the ratio of the misclassification costs might vary. In Figure 7 for a 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, A should be compared to the trivial classifier marked C on the solid curve at the origin.

The dashed and solid cost curves in Figure 8 correspond to the dashed and solid ROC curves in Figure 7. The horizontal line atop the solid cost curve corresponds to classifier B[2]. The vertical distance between the 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.

7. Averaging Multiple Curves

The dashed lines Figure 9 are two ROC curves. If these are the result of learning from 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 (1982) suggest two methods, pooling and "averaging", and Provost et al. (1998) propose an alternative averaging method.

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 9, with each vertex corresponding to a vertex from one or other of the dashed curves. Figure 10 shows the equivalent two cost curves 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

2. It is horizontal because $FP = 1 - TP$ for this classifier.

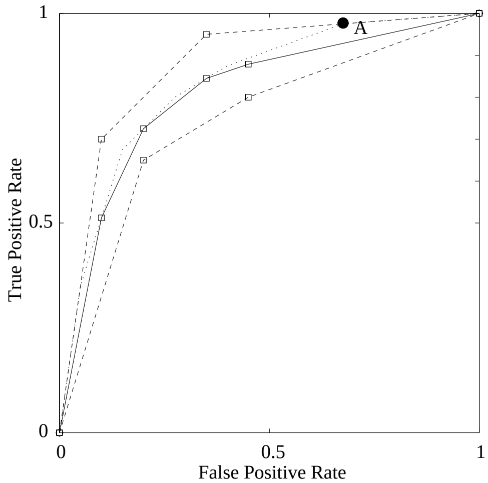

Figure 9: Average ROC Curves

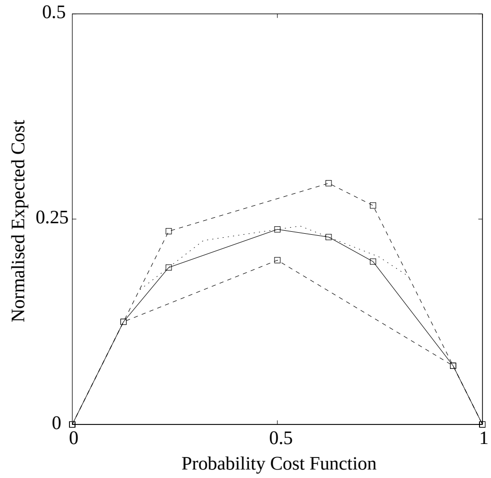

Figure 10: Average Cost Curves

to cost space the dotted line results. Similarly, the dotted line in Figure 9 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, under the Provost et al. method of ROC averaging, if they have the same FP value. Pooling, or other methods of averaging ROC curves, will all produce different results because they put the points on the two curves into correspondence in different ways.

For the cost curves points correspond if they have the same $PCF(+)$ value. It is illuminating to look at the dotted line in the top right hand corner of Figure 9. 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 curve average has a very clear meaning: at each operating point it gives the average normalised expected cost for that operating point.

8. Confidence Intervals on Costs

The measure of classifier performance is derived from a confusion matrix produced from some sample of the data. As there is likely to be variation between samples, the measure is, itself, a random variable. So some estimate of its variance is useful, which usually takes the form of a confidence interval. The most common approach to producing a confidence interval is to assume that the distribution of the estimate belongs to, or is closely approximated by, some parametric family such as Gaussian or Student-t. An alternative, data driven, method has become popular in recent times which does not make any parametric assumptions. Margineantu and Dietterich (2000) described how one such non-parametric approach called the bootstrap (Efron & Tibshirani, 1993) can be used to generate confidence intervals for predefined cost values. We use a similar technique, but for the complete range of class distributions and misclassification costs.

The bootstrap method is based on the idea that new samples generated from the available data are related to that data in the same way that the available data relates to the original population. Thus the variance of an estimate based on the new samples should be a good approximation to its true variance. Confidence limits are produced by resampling from the original matrix to create nmerous new confusion matrices of the same size. The exact way bootstrapping is carried out depends on the sampling scheme. We propose a resampling method analogous to stratified cross validation, in which the class frequency is guaranteed to be identical in every sample.

Pred. / Act.	Pos	Neg	
Pos	16 P1	4 1−P1	20 m
Neg	4 P2	6 1−P2	10 n

Figure 11: Binomial Sampling

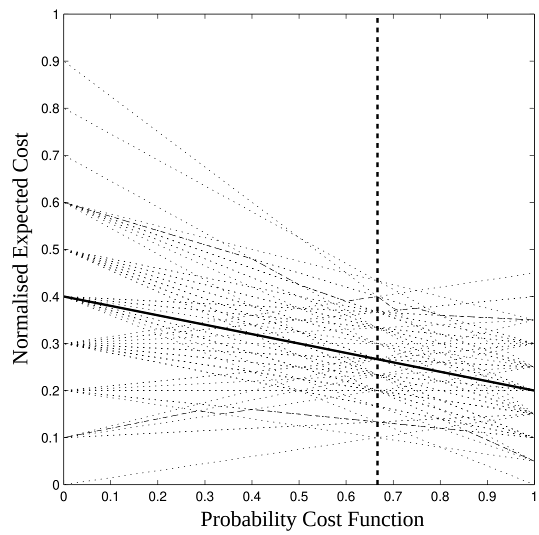

Figure 12: Binomial Intervals

For example, consider the confusion matrix of Figure 11. There are 30 instances, 20 of which are positive and 10 negative. The classifier correctly labels 16 out of 20 of the positive class, but only 6 out of 10 of the negative class. We fix the row totals at 20 and 10, and treat the two rows as independent binomial distributions with probabilities $P1 = 16/20 = 0.8$ and $P2 = 4/10 = 0.4$, respectively, of assigning a positive label to an example.

A new matrix is produced by randomly sampling according to these two binomial distributions until the number of positive and negative instances equal the corresponding row totals. For each new confusion matrix, a dotted line is plotted in Figure 12 representing the new estimate of classifier performance. For ease of exposition, we generated 100 new confusion matrices (typically at least 500 are used for an accurate estimate of variance). To find the 90% confidence limits, if we had values just for one specific x-value, the fifth lowest and fifth highest value could be found. This process is repeated for each small increment in the PCF(+) value. These upper and lower confidence limits are shown as the dashed lines in Figure 12. The bold line represents the performance of the classifier based on the original confusion matrix.

9. Testing if Performance Differences are Significant

The difference in performance of two classifiers is statistically significant if the confidence interval around the difference does not contain zero. The method presented in the previous section can be extended to do this, by resampling the confusion matrices of the two classifiers simultaneously, taking into account the correlation between the two classifiers. A single resampling thus produces a pair of confusion matrices, one for each classifier, and therefore two lines in cost space. However, instead of plotting the two lines, we plot the difference between the two lines (which is itself a line). We can repeat this process a large number of times to get a large number of lines and then, as above, extract a 90% confidence interval from this set of lines. This is the confidence interval around the difference between the classifiers' performances.

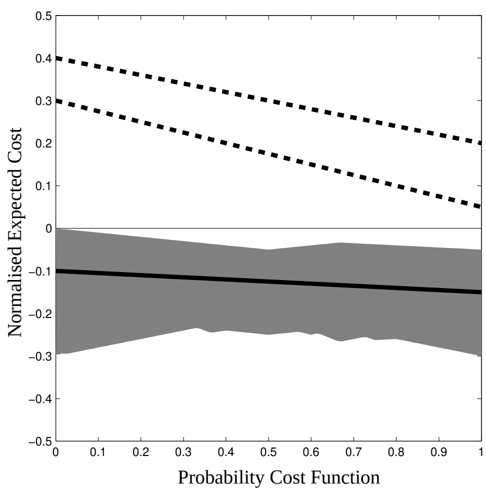

Figure 13: CI for the Difference

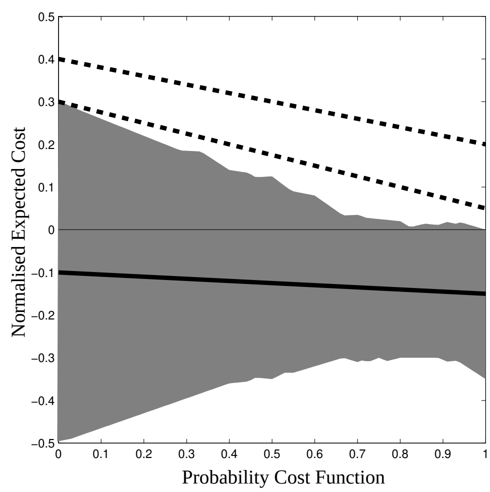

Figure 14: CI for Reduced Correlation

The thick continuous line at the bottom of Figure 13 represents the mean difference between performance of the two classifiers (which are shown in the figure as bold dashed lines). The shaded area represents the confidence interval of the difference, calculated as just described. As the difference can range from ± 1 the y-axis has been extended. Here we see that the confidence interval does not contain zero, so the difference between the classifiers is statistically significant. Figure 14 shows the same two classifiers but with their classifications less correlated. Notably, the confidence interval is much wider and includes zero, so the difference is not statistically significant. Thus cost curves give a nice visual representation of the difference in expected cost between two classifiers across the full range of misclassification costs and class frequencies. The cost curve representation also makes it clear that performance differences might be signficant for some range of operating points but not others.

10. Conclusions

This paper has demonstrated shortcomings of ROC curves for visualizing classifier performance, and showed that cost curves overcome these problems. We do not, however, contend that cost curves are always better than ROC curves. For example, for visualizing the work-force utilization measure of performance, ROC curves are distinctly superior to cost curves. But for many common visualization requirements, cost curves are by far the best alternative and we recommend their routine use instead of ROC curves for these purposes.

11. Acknowledgements

We would like to acknowledge Alberta Ingenuity Fund for its support of this research through the funding of the Alberta Ingenuity Centre for Machine Learning.

References

Drummond, C., & Holte, R. C. (2000). Explicitly representing expected cost: An alternative to roc representation. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 198–207.

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and scene analysis.* Wiley, New York.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

Margineantu, D. D., & Dietterich, T. G. (2000). Bootstrap methods for the cost-sensitive evaluation of classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 582–590.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 217–225 San Francisco. Morgan Kaufmann.

Provost, F., & Fawcett, T. (1998). Robust classification systems for imprecise environments. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 706–713 Menlo Park, CA. AAAI Press.

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 San Francisco. Morgan Kaufmann.

Swets, J. A. (1967). *Information Retrieval Systems.* Bolt, Beranek and Newman, Cambridge, Massachusetts.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems : methods from signal detection theory.* Academic Press, New York.

Van Rijsbergen, C. J. (1979). *Information retrieval.* Butterworths, London.

Webb, G., & Ting, K. M. (2004). On the application of ROC analysis to predict classification performance under varying class distributions.. *Machine Learning* (to appear).