

Have we overestimated the value of ROC analysis under varying class distributions?

Geoffrey I. Webb

Kai Ming Ting

*School of Computer Science and Software Engineering and
Gippsland School of Computing and Information Technology,
Building 26, Monash University*

Victoria, 3800, Australia

tel: +61 3 99053296

fax: +61 3 99055146

webb@infotech.monash.edu.au

Abstract. We counsel caution in the application of ROC analysis for prediction of classifier accuracy under varying class distributions. We argue that it is only reasonable to expect ROC analysis to provide accurate error prediction under varying class distributions if the attributes are causally dependent upon the classes and the classes do not contain causally relevant subclasses whose frequency may vary at different rates.

Keywords: Model evaluation, ROC analysis

1. Introduction

ROC analysis has appeared to offer more robust evaluation of the relative prediction performance of alternative models than traditional comparison on the basis of relative error (Weinstein and Fineberg, 1980; Provost, Fawcett and Kohavi, 1998; Adams and Hand, 1999; Duda, Hart and Stork, 2001). Rather than considering raw error, ROC analysis decomposes performance into true and false positive rates. Different ROC profiles will be more or less desirable under different class distributions and different error cost functions. This analysis is held to provide more robust comparative evaluation of expected performance on future test data than simple comparison of error, which assumes the observed class distribution and the all errors have equal cost. While we do not question the ability of ROC analysis to provide useful comparative evaluation in a context where the distribution of attributes and classes will remain constant but the relative error costs are not, we believe that its value in the context of test data with unknown class distributions has been overestimated. At face value, the following quotations assert that ROC techniques provide a class frequency independent measure. In the remainder of this paper we argue that this is only true under very specific constraints.

[ROC] is the only measure available that is uninfluenced by decision biases and prior probabilities ... (Swets, 1988)

The Area Under the ROC Curve ... is invariant to prior class probabilities ... (Bradley, 1997)

ROC graphs illustrate the behaviour of a classifier without regard to class distribution or error cost (Provost and Fawcett, 2001).

Our concern with these assertions relates to the claims about class distributions or prior probabilities. We assume that we are talking about evaluating the expected performance of a model on future data by assessing its performance on sample data. We will denote probabilities in the sample distribution as $P(\cdot | \textit{train})$ and probabilities in the future data to which the model is to be applied as $P(\cdot | \textit{test})$. We assume a model mapping instances from a description space X to a description space Y^1 . Y is the set of *classes* $\{p, n\}$, where p is the positive class and n is the negative class. Each $x \in X$ is a vector of attribute values $x = \langle x_1, \dots, x_k \rangle$.

Swets (1988) appears to be asserting that ROC analysis provides a useful assessment of the expected performance of a model irrespective of whether $P(Y = p | \textit{test}) = P(Y = p)$ or not. Bradley (1997) and Provost and Fawcett (2001) appear to be making a slightly different assertion, that ROC analysis provides a useful assessment of the expected performance of a model irrespective of whether $P(Y = p | \textit{test}) = P(Y = p | \textit{train})$. We address our comments to the latter assertion, as we believe that this is the most relevant in the machine learning context. However, closely related arguments are equally applicable to the former assertion.

To summarize the basis for our misgivings, a ROC curve will only capture error performance on test data under varying class distributions if its estimates of the true and false positive rates hold for the test data. As we show, this condition is only satisfied under a number of very strong constraints. In consequence, we argue that the application of ROC analysis for comparative evaluation of learning algorithms should be restricted to contexts in which it is desired to compare performance under unknown error costs or for which it is desired to compare performance under unknown class distributions when it is reasonable to assume that true and false positive rates will remain invariant or only vary slightly. We argue further that it is only plausible to expect this latter condition to hold when all the attribute values of X that are used by the model are causally related to Y such that the X values depend

¹ We deliberately avoid describing XY as a joint distribution as the claims that we discuss relate to outcomes under varying distributions and describing XY as a distribution adds potential confusion as to which distribution in particular it refers.

on the Y values. There is a further constraint that there must not be causally relevant sub-categories of the Y values whose distributions may vary at different rates.

In the remainder of this paper we describe ROC analysis, provide a trivial example of how changes to data distributions that cause the class distribution to change may also change the true and false positive rates, discuss the circumstances under which the class distribution may change while the true and false positive rates do not, and then conclude with a brief discussion.

2. ROC analysis

We assume that ROC analysis is used to assess the expected performance of a model $M(X) \rightarrow Y$, a function from X s to Y s. We assume that there is a target relationship $T(X, Y)$ that describes the true relationship between X and Y . We envisage this relationship in terms of a Bayesian network, but allow that it might be captured by any equivalent formalism. We assume that this target relationship does not vary from the training to the test data. In other words, we assume that there is no concept drift. By this we mean that the prior probabilities associated with nodes with no parents may change but that the conditional probabilities that capture direct inter-relationships between attributes may not. Concept drift is related to the problem we address herein. However, we assume that it is clear to all concerned that it is unreasonable to expect any evaluation method to provide accurate predictions about the future performance of a model in the presence of concept drift and in the absence of information about the nature of that concept drift. Rather, we assume that those proposing the use of ROC analysis to predict expected performance under varying class distributions are assuming that there is a change in the composition of the data from training to test data but that change does not alter the target concept.

ROC analysis is used to evaluate expected performance of M when it is applied to previously unseen data that we call the *test data*. This evaluation is achieved by analysis of the performance of M when applied to the training data.

The analysis is based on observation of four types of outcome, where we introduce the notation $O(x)$ to denote the observed value of Y for an observed object with $X = x$:

- *true positives*, where $M(x)=p \wedge O(x)=p$,
- *false positives*, where $M(x)=p \wedge O(x)=n$,

- *true negatives*, where $M(x)=n \wedge O(x)=n$,
- *false negatives*, where $M(x)=n \wedge O(x)=p$,

The true positive and false positive rates of a classifier are defined as follows.

$$\begin{aligned}
 TP &= \frac{\textit{positives correctly classified}}{\textit{total positives}} \\
 &= \frac{P(M(X)=p \wedge O(X)=p | \cdot)}{P(O(X)=p | \cdot)} \\
 &= P(M(X)=p | O(X)=p \wedge \cdot) \\
 &= P(M(X)=p | Y=p \wedge \cdot).
 \end{aligned}$$

$$\begin{aligned}
 FP &= \frac{\textit{negatives incorrectly classified}}{\textit{total negatives}} \\
 &= \frac{P(M(X)=p \wedge O(X)=n | \cdot)}{P(O(X)=n | \cdot)} \\
 &= P(M(X)=p | O(X)=n \wedge \cdot) \\
 &= P(M(X)=p | Y=n \wedge \cdot).
 \end{aligned}$$

We make explicit mention of a context (\cdot) for these probability assessments to remind the reader that these quantities can only be measured given a reference distribution.

ROC space is defined as a coordinate system. The y-axis represents TP and the x-axis represents FP . The performance of a classifier is represented as a point in this space, denoted as a (FP, TP) -pair. For a model that produces a continuous output, such as a posterior probability, a series of (FP, TP) -pairs can be obtained by varying the decision threshold at which a positive class prediction is made. The resulting curve of (FP, TP) -pairs is called the ROC curve, originating from $(0,0)$ and ending at $(1,1)$.

We address here the adequacy of any of these points as a measure of expected performance under varying class distributions, that is, the ROC curve assessment for any given decision threshold. Our model M can be considered the model formed by a classifier under any one of its decision thresholds or as a classifier that does not admit to multiple decision thresholds.

ROC assessment of error under varying class distributions relies on the assumption that

$$\textit{error} = (1.0 - TP) \times P(Y=p | \textit{test}) + FP \times P(Y=n | \textit{test}), \quad (1)$$

where TP and FP are the true and false positives on the training data distribution, but error may be error under any class distribution.

Under this assumption, each (FP, TP) -pair defines error rate irrespective of class distribution. However, this assumption only holds under very limited conditions, conditions that we believe are often violated in real-world training and testing scenarios.

We presume that (1) rests upon a further assumption that TP and FP will remain invariant from the training to the test data. It is true that (1) can hold under propitious circumstances where TP and FP vary from the training to the test data in such a way that an increase in TP , for example, is matched by exactly the right increase in FP for the resulting application of (1) to derive the true error on the test data. But it appears unlikely that this is the rationale for belief in (1).

For TP to remain invariant while $P(Y = p | \cdot)$ varies requires that $P(M(X) = p | Y = p \wedge \cdot)$ remain invariant. Likewise, for FP to remain invariant while $P(Y = n | \cdot)$ varies requires that $P(M(X) = p | Y = n \wedge \cdot)$ remain invariant. We can expect these invariances if the process that generates the training and testing distributions results from a systematic manipulation of the class. We cannot, in general, expect it if the difference in distributions results from a systematic manipulation of the attribute values without reference to the class.

3. An example

We provide an extremely simple example to illustrate how alterations to the distribution of the attributes without regard to the distribution of the class may both alter the distribution of the class and violate (1). Consider a learning task inspired by Quinlan's (1987) classic example of deciding whether to play golf. There are two attributes, *Playing Conditions*, with the two values *Pleasant* and *Unpleasant*, and *Other Commitments* with the two values *Busy* and *Free*. The classes are *Play* and *Don't Play*, with the former considered the positive class. The target concept is *Play* if and only if *Pleasant* and *Free*. As we rule out concept drift from consideration, we do not allow this concept to alter. To make the example as simple as possible, we assume that the attributes are independent of each other. Our ability to construct such an example in no way depends upon this assumption. This assumption is made solely as a matter of convenience. The training data (or at least the data from which the ROC curve is to be derived) is taken from observations drawn over a year for which the frequencies of each of the four combinations of attribute values are equal. Table I displays the four combinations of independent variable together with

Table I. Example data distributions

| Object | Initial | Retire | Inter- mediate | Propitious | Paradise |
|------------------------------|---------|--------|-------------------|------------|----------|
| Pleasant, Free, Play | 0.25 | 0.50 | 0.50 | 0.50 | 0.50 |
| Pleasant, Busy, Don't Play | 0.25 | 0.00 | 0.21 | 0.17 | 0.50 |
| Unpleasant, Free, Don't Play | 0.25 | 0.50 | 0.21 | 0.25 | 0.00 |
| Unpleasant, Busy, Don't Play | 0.25 | 0.00 | 0.08 | 0.08 | 0.00 |

the associated class. The column titled *Initial* shows the frequency with which each combination appears in the training data. To remove sampling error as an issue, we assume that the sample frequencies exactly match the probabilities.

In order to cast light on ROC analysis we require a model to analyze. In order to demonstrate our point, in the case where the class is uniquely determined by the attribute values, we require only that there be two combinations of attribute values x^1 and x^2 such that $O(x^1) = O(x^2)$ and $M(x^1) \neq M(x^2)$, or in other words, that at least one class is sometimes, but not always, misclassified. If these minimal conditions are not satisfied, *TP* and *FP* must be invariant no matter what the data distribution. An extremely simple model is required to satisfy this constraint for this extremely simple example. To illustrate the same points with respect to more complex models it would be necessary to formulate a more complex example. Assume we apply decision stump learning (Holte, 1993). We might form a model that classifies an object as *Play* if and only if *Pleasant*. For this model, the true positive ratio is 1.0 (all *Play* objects are correctly labelled) and the false positive ratio is 1/3 (pleasant but busy days are misclassified).

Suppose now we move to evaluation data for which there is a different class distribution to that of the training data. ROC analysis is supposed to apply irrespective of the class distribution. For the sake of illustration we will increase the frequency of *Play* in the test data to 0.5. Note, however, that the particular frequency is not important to our example. The same effect will be apparent for any change in the class distribution. All that alters with different distributions is the magnitude of the effect.

Recall that $P(\textit{Pleasant})$ and $P(\textit{Free})$ are independent. We assume that this property also holds of $P(\textit{Pleasant}|\textit{test})$ and $P(\textit{Free}|\textit{test})$. As we are increasing the frequency of *Play* to 0.5 we require that $P(\textit{Pleasant}\&\textit{Free}|\textit{test}) = 0.5$. As $P(\textit{Pleasant}|\textit{test})$ and $P(\textit{Free}|\textit{test})$ are independent, it follows that we require that $P(\textit{Pleasant}|\textit{test}) \times P(\textit{Free}|\textit{test}) = 0.5$. Four of the infinite number

of combinations of $P(\textit{Pleasant} | \textit{test})$ and $P(\textit{Free} | \textit{test})$ for which the desired class distribution are obtained are:

1. $P(\textit{Pleasant} | \textit{test})$ remains 0.5 while $P(\textit{Free} | \textit{test})$ rises to 1.0 (we retire!), illustrated in the *Retire* column of Table I;
2. $P(\textit{Pleasant} | \textit{test})$ and $P(\textit{Free} | \textit{test})$ both rise to 0.72, illustrated in the *Intermediate* column of Table I;
3. $P(\textit{Pleasant} | \textit{test})$ rises to 0.67 and $P(\textit{Free} | \textit{test})$ rises to 0.75, illustrated in the *Propitious* column of Table I; and
4. $P(\textit{Pleasant} | \textit{test})$ rises to 1.0 while $P(\textit{Free} | \textit{test})$ remains 0.5 (we move to paradise!), illustrated in the *Paradise* column of Table I.

These are only four out of an infinite number of possible combinations of values of $P(\textit{Pleasant} | \textit{test})$ and $P(\textit{Free} | \textit{test})$ for which $P(\textit{Play} | \textit{test}) = 0.5$. For all alternatives the true positive ratio will remain 1.0. This is because our model happens to be an overgeneralization of the true concept. However, of all the infinite number of combinations of $P(\textit{Pleasant} | \textit{test})$ and $P(\textit{Free} | \textit{test})$ for which the new class distribution are obtained, only for exactly those propitious values $P(\textit{Pleasant} | \textit{test}) = 2/3$ and $P(\textit{Free} | \textit{test}) = 0.75$ does the false positive ratio remain at $1/3$.

For the retirement scenario, the false positive ratio becomes 0.0 and the ROC analysis will overestimate error. For the intermediate scenario, the false positive ratio rises to 0.41^2 and the ROC analysis will underestimate error. For the paradise scenario, the false positive ratio becomes 1.0 and the ROC analysis will again underestimate error. If the true or false positive rates do change then ROC analysis' prediction of the error rate under a new class distribution will be incorrect.

4. How likely is it that *TP* and *FP* will remain invariant?

For ROC analysis to provide information about the error rates that may be expected under varying class distributions, the true and false positive rates must remain invariant across changes in class distribution. As our simple example has shown, even for a trivial concept, it takes very precise manipulation of the frequency of the independent variables to change the class distribution without also changing the true and false

² This is calculated using intermediate values of greater precision than those displayed in Table I. The true value of $P(\textit{Pleasant})$ such the $P(\textit{Pleasant}) = P(\textit{Free})$ and $P(\textit{Pleasant}) \times P(\textit{Free}) = 0.5$ is $P(\textit{Pleasant}) = \sqrt{0.5}$.

positive rates of a simple model. It appears extremely implausible to expect that changes in the class frequency under real-world conditions should normally be accompanied by invariant true and false positive rates for arbitrary models.

The only normal circumstance under which it seems reasonable to expect class frequency to change while true and false positive rates remain invariant (or, at least, to vary only insofar as is due to sampling error) is when the probabilities of the values of the attributes are determined by the value of the class rather than vice versa. That is, when $P(m(X)|Y \wedge \cdot)$ for the distribution from which the sample is drawn remains invariant across varying class distributions. We use $m(X)$ to denote the subset of X to which the model M refers.

One example of when this will be the case is when data is drawn from a single database using random sampling for which the probability of an object's selection depends solely upon its class. Such stratified sampling may well occur during machine learning experiments. However, it is difficult to conceive of a real-world application where a change in class distribution through some natural process replicates the effect of stratified sampling.

The only other reason that we might expect $P(m(X)|Y \wedge \cdot)$ to remain invariant is if $m(X)$ depends on Y . This is at least superficially credible in some circumstances. For example, for medical diagnosis it is credible that X will consist of signs and symptoms caused by the disease represented by $Y = p$. During fraud detection, it is credible that $m(X)$ will be a signature of fraudulent behavior that is caused by the presence of the fraud represented by $Y = p$.

There are many other circumstances where it is not even superficially credible, however. If X represents the operating characteristics of a production line and $Y = p$ represents a fault in a product manufactured under conditions X , it is not credible that the production of a faulty product caused the production line to be in a specific configuration. Rather, the configuration causes the fault. A change in the frequency of faults will result from a change in the frequency of specific configurations, and ROC analysis will fail to predict the rate of faults under the new class distribution. Similar examples hold for predicting propensity to purchase from socio-economic data, scholastic achievement from entry scores, or indeed, any task where (any of) the X values represent measures taken before the Y values are determined, as most views of causation require that the cause precedes the effect (Sosa and Tooley, 1994, for example).

But even for the circumstance where there is a causal relationship from Y to X , $P(m(X)|Y \wedge \cdot)$ may still vary from training to test set. If $Y = p$ represents a superclass of related subclasses, such as any of a

number of types of hypothyroid disease or any of a number of types of fraud, a change in the frequency of $Y = p$ is likely to represent differing degrees of change in the frequency of each of the subclasses. Suppose we have just two subclasses, a and b , and each has a signature set of X values, $s_a(X)$ and $s_b(X)$. The model is thus $Y = p$ if and only if $s_a(X)$ or $s_b(X)$. If the frequency of $Y = p$ increases due to an increase in the frequency of a but the frequency of b remains unchanged then $P(m(X) | Y \wedge \cdot)$ will change and ROC analysis can be expected to fail to accurately predict model performance.

Alternatively, even if the target class is uniform with a single signature, suppose that X contains attributes that do not relate to Y , but that the learning system incorrectly incorporates one such value into its model. Now, an increase in the frequency of $Y = p$ will increase the frequency of the signature in X , but it cannot be expected to affect the frequency of the spurious attribute incorporated in the model, and again $P(m(X) | Y \wedge \cdot)$ can be expected to vary from training to test data.

As a final scenario, consider the circumstance where X contains both attributes A caused by Y , and attributes B that cause Y . For example, the target disease, or fraud, might be more prevalent among a specific age group. In this case a good learning system should incorporate both attributes A and B in its model. Between the formation of the model and its application the frequency of B alters (the population ages or a company sets out to acquire customers in a particular age group). Again, $P(m(X) | Y \wedge \cdot)$ will change from the training to the test data and ROC analysis can be expected to fail to accurately predict model performance.

5. Conclusions

We have argued that literal interpretation of a number of statements in the literature about ROC analysis might lead some practitioners to over-estimate its capacity to predict model error under varying class distributions. We have provided a detailed example that illustrates our concern. We have argued that it is only realistic to expect ROC analysis to accurately predict model error when all of the following are satisfied:

- all the attributes used by the model depend upon the class, rather than the reverse or being independent, and
- the class does not contain causally relevant sub-populations whose frequency varies at different rates.

A further constraint on the likely accuracy of ROC analysis under varying class distributions is one that we have deliberately set to one side until this point: the possibility that the variation might result from concept drift. It is surely incumbent upon a practitioner intending to employ ROC analysis to predict error under varying class distributions to assess first whether any of these possible confounds is present.

In this paper we have concentrated on the issue of the circumstances under which a single ROC point can be expected to provide reliable prediction of expected error under variations in class distribution. We believe that we have established that a number of very specific constraints must be satisfied in order for such evaluation to be accurate. This conclusion extends directly to the issues of whether ROC curves and the area under the ROC curve provide useful measures of expected comparative performance under varying class distributions. This is because the ROC curve is derived from a sequence of ROC points. If each of the points does not provide a reliable estimate of prediction performance then it follows that neither the curve nor the area under the curve provides a reliable estimate either.

Our investigations lead naturally to the question, are there alternatives to ROC analysis? We believe that there are. It is clearly not possible to provide an area-under-the-ROC-curve style assessment of comparative performance under the possibility of unconstrained variations in any aspect of the data distribution. This follows from a straightforward no-free-lunch (Wolpert, 1992) style argument. The average error performance of any two models over the space of all possible data distributions will be identical. Comparative area-under-the-ROC-curve style assessment could be achieved for constrained variations in data distributions by performing Monte Carlo simulations that compare performance on random data sets drawn from the space of data distributions under consideration.

There exists a very straightforward and simple manner in which to evaluate expected error from a specific model given a specific variation in class distribution. All that is required is to form a test set drawn from the new distribution and assess the model's performance thereon. It might be argued that such an approach is restrictive, because it requires labelled data. However, unless such labelled data is available it is surely not possible to determine the new class frequencies and hence not possible to estimate error using ROC techniques anyway. Forming a test set from the new data distribution can be expected to provide reasonable estimates of expected error under any form of change in data distribution, including concept drift. Hence, we recommend the use of this approach as it is subject to none of the constraints that

limit the potential applicability of ROC analysis to predict error under variations in class distribution.

References

- Adams, N. M. and Hand, D.J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32, 1139–1147.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159.
- Duda, O.R., Hart, P.E. and Stork, D.G. (2001). *Pattern Classification*, John Wiley.
- Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1), 63–90.
- Provost, F. and Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning*, 42, 203–231.
- Provost, F., Fawcett, T. and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of The Fifteenth International Conference on Machine Learning*, 43–48. San Francisco: Morgan Kaufmann.
- Quinlan, J. R. (1987). Learning decision trees. *Machine Learning*, 1(1), 1–25.
- Sosa, E. and Tooley, M. Editors. (1994). *Causation*, Oxford University Press.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240, 1285–1293.
- Weinstein, M.C. and Fineberg, H.V. (1980). *Clinical Decision Analysis*, Saunders.
- Wolpert, D. H. (1992). On the connection between in-sample testing and generalization error. *Complex Systems*, 6, 47–94.

