On the application of ROC analysis to predict classification performance 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 75, 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 performance under varying class distributions. We argue that it is not reasonable to expect ROC analysis to provide accurate prediction of model performance under varying distributions if the classes contain causally relevant subclasses whose frequencies may vary at different rates or if there are attributes upon which the classes are causally dependent.

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 of relative error (Weinstein and Fineberg, 1980; Bradley, 1997; 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 target data than simple comparison of error, which assumes the observed class distribution and does not reflect any differences in the cost of different types of error. While we do not question the use of ROC analysis for comparative evaluation across the space of all possible cost functions, we are concerned that the literature has failed to counsel sufficient caution about its use in the context of changing class distributions. The literature contains a number of statements, such as the following, suggesting that ROC analysis might be of value for evaluating expected classifier performance under varying class distributions, but other than the implicit qualification of the third quote, none to our

Prepublication draft of paper accepted for publication in *Machine Learning*

knowledge are accompanied by warnings that in some contexts this might be inappropriate.

> [ROC] is the only measure available that is uninfluenced by decision biases and prior probabilities ... (Swets, 1988)
>
> ROC curves describe the predictive behavior of a classifier independent of class distributions or error costs, so they decouple classification performance from these factors. (Provost, Fawcett and Kohavi, 1998)
>
> ... the key assumption of ROC analysis is that true and false positive rates describe the performance of the model independently of the class distribution ... (Flach, 2003)

Note that the application of ROC analysis to evaluate classifier performance under varying cost functions does not require the assumption stated in the third quote. This assumption is only required if ROC analysis is to be employed to predict performance under varying class distributions. We argue that it is not always reasonable to make such an assumption. Specifically, this assumption only appears reasonable if it is expected that the true and false positive rates will remain invariant while the class distribution changes, an assumption that does not appear sensible to make without justification.

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 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 $\lambda(\mathcal{X}) \rightarrow \mathcal{Y}$, a function from a description space $\mathcal{X}$ to a description space $\mathcal{Y} = \{p, n\}$, where p is the positive class and n is the negative class. Each $x \in \mathcal{X}$ is a vector of attribute values $x = \langle x_1, \ldots x_k \rangle$. We use uppercase letters $A \ldots Z$ to represent random variables and lowercase letters $a \ldots z$ to represent values. X is reserved for a random variable over the values of a description space $\mathcal{X}$ and Y for a random variable over the values of a description space $\mathcal{Y}$. Much of the discussion relates to probability distributions. For ease of exposition we restrict consideration to probability distributions over discrete valued variables, although the principles extend directly to the case of probability density functions over continuous variables. For

notational convenience, the variable is omitted in a probability term in which it is clearly implied by the context, for example $P(a \mid b)$ stands for $P(A{=}a \mid B{=}b)$. We use the notational shorthand of a probability term with values omitted to represent the distribution of the term over all values of the variable(s). For example, $P(X \mid Y)$ represents the distribution of $P(X{=}x \mid Y{=}y)$ over all $x \in X$, $y \in Y$.

We address the situation where ROC analysis is used to predict the performance of λ when it is applied to previously unseen data that we call *target* data. This evaluation is achieved by analysis of the performance of λ when applied to another set of data that we call the *base* data.

The analysis is based on observation of four types of outcome:

- *true positives*, where $\lambda(X){=}p \wedge Y{=}p$,
- *false positives*, where $\lambda(X){=}p \wedge Y{=}n$,
- *true negatives*, where $\lambda(X){=}n \wedge Y{=}n$,
- *false negatives*, where $\lambda(X){=}n \wedge Y{=}p$,

The true positive and false positive rates are defined as follows.

$$\begin{aligned} TP(\cdot) &= \frac{\textit{true positives in context } \cdot}{\textit{total positives in context } \cdot} \\ &= P(\lambda(X){=}p \mid Y{=}p \wedge \cdot). \end{aligned}$$

$$\begin{aligned} FP(\cdot) &= \frac{\textit{false positives in context } \cdot}{\textit{total negatives in context } \cdot} \\ &= P(\lambda(X){=}p \mid Y{=}n \wedge \cdot). \end{aligned}$$

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. We will use $TP(target)$, $FP(target)$ and $P(A \mid target)$ to denote each of the relevant functions evaluated on the distribution with respect to which we are interested in predicting classification performance. We use $TP(base)$, $FP(base)$ and $P(A \mid base)$ to denote each of the relevant functions evaluated on base data by means of which it is intended to predict $TP(target)$ and $FP(target)$.

ROC space is defined as a coordinate system. The y-axis represents $TP(base)$ and the x-axis represents $FP(base)$. 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 λ 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 classification performance under varying class distributions relies on the assumption that a model's $TP(\cdot)$ and $FP(\cdot)$ rates will remain invariant as the class distribution changes. Under this assumption, each (FP,TP)-pair defines expected performance irrespective of class distribution. However, this assumption only holds under specific conditions, conditions that we believe are often violated in real-world scenarios.

For $TP(\cdot)$ to remain invariant while $P(Y{=}p \mid \cdot)$ varies requires that $P(\lambda(X){=}p \mid Y{=}p \wedge \cdot)$ remain invariant. Likewise, for $FP(\cdot)$ to remain invariant while $P(Y{=}n \mid \cdot)$ varies requires that $P(\lambda(X){=}p \mid Y{=}n \wedge \cdot)$ remain invariant. We can expect these invariances if the process that generates the base and target 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 a 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 alter true and false positive rates. Consider a learning task inspired by Quinlan's (1987) classic example of deciding whether to play golf. We seek to predict the behavior of a golf enthusiast called John. John's behavior can be accurately predicted with reference to 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*, representing respectively whether John plays golf or does not, with the former considered the positive class. The underlying concept is *Play* if and only if *Pleasant* and *Free*. That is, John plays golf whenever the weather is pleasant and he has no other commitments. As we do not consider concept drift, we do not

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

allow this concept to alter. To make the example as simple as possible, we assume that the attributes are independent of each other. That is, John's commitments do not affect the weather and the weather does not affect John's commitments. Our ability to construct an example in no way depends upon this simplifying assumption, however. The base data are 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 X values together with the associated class. The column titled *Initial* shows the frequency with which each combination appears in the base data. To remove sampling error as an issue, we assume that the sample frequencies exactly match the true 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 at least one class is sometimes, but not always, misclassified. If these minimal conditions are not satisfied, $TP(\cdot)$ and $FP(\cdot)$ must be invariant no matter what the data distribution. Assume we apply decision stump learning (Holte, 1993). We might form a model that classifies an occassion as $Play$ if and only if $Pleasant$. For this model, $TP(base) = 1.0$ (all Play objects are correctly labelled) and $FP(base) = 1/3$ (pleasant but busy days are misclassified).

Suppose now we move to target data for which there is a different class distribution from that of the base 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 target data to 0.5. Note, however, that this 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. Note also that we are addressing here the situation where the change from the base to the target data represents a change in the underlying distributions from which the data are drawn, rather than a change in the way in which the data are sampled.

As we are increasing the frequency of $Play$ to 0.5 we require that $P(Pleasant \wedge Free \mid target) = 0.5$. As $Pleasant$ and $Free$ are independent, it follows that we require that $P(Pleasant \mid target) \times P(Free \mid target) = 0.5$. That is, because the weather and John's commitments determine the value of the class variable, it is only by varying both the weather and John's commitments precisely in conjunction that it is possible to obtain a particular class distribution. Four of the infinite number of combinations of $P(Pleasant \mid target)$ and $P(Free \mid target)$ for which the desired class distribution are obtained are:

1. $P(Pleasant \mid target)$ remains 0.5 while $P(Free \mid target)$ rises to 1.0 (John retires!), illustrated in the *Retire* column of Table I;

2. $P(Pleasant \mid target)$ and $P(Free \mid target)$ both rise to 0.71, illustrated in the *Intermediate* column of Table I;

3. $P(Pleasant \mid target)$ rises to 0.67 and $P(Free \mid target)$ rises to 0.75, illustrated in the *Propitious* column of Table I; and

4. $P(Pleasant \mid target)$ rises to 1.0 while $P(Free \mid target)$ remains 0.5 (John moves to paradise!), illustrated in the *Paradise* column of Table I.

For all alternatives the true positive rate 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 $P(Pleasant \mid target)$ and $P(Free \mid target)$ for which the new class distribution are obtained, only for exactly those propitious values $P(Pleasant \mid target) = 2/3$ and $P(Free \mid target) = 0.75$ does the false positive rate remain at 1/3. For this example, for ROC analysis to successfully predict classification performance under a change of class distribution requires that the world is organized so that a golfer's commitments can only (or are most likely) to change only in conjunction with specific changes in the weather.

If John ceases having other commitments but the weather does not change (the retirement scenario), the false positive rate becomes 0.0 and the ROC analysis will overestimate it. For the intermediate scenario in which there are equal increases in the frequencies both of pleasant weather and of John having no commitments, the false positive rate rises to 0.41[1] and the ROC analysis will underestimate it. If the weather improves so as to be invariably pleasant but John's commitments do

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

not change (the paradise scenario), the false positive rate becomes 1.0 and the ROC analysis will again underestimate it.

4. Discussion

For ROC analysis to provide information about the performance 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 precise manipulation of the frequency of the X values to change the class distribution without also changing the true and false positive rates of a simple model in which the class is causally dependent upon X. Rather, it is only reasonable to expect true and false positive rates to remain invariant when $P(\pi_\lambda(X) \mid Y \wedge \cdot)$ remains invariant across varying class distributions. We use $\pi_\lambda(X)$ to denote the projection of the value of X onto the attributes to which the model $\lambda(X)$ refers. This is required in order to allow for the possibility that a given model may ignore some attributes and hence that changes in the conditional probabilities relating to those attributes will not affect directly the performance of the model.

This will be the case when base data are sampled using random sampling for which the probability of an object's selection depends solely upon its class. Such stratified sampling may well occur during data mining projects and indeed over-sampling of infrequent classes can be a valuable learning technique. In such a context, ROC analysis may be used to recover, from the performance of a model on the stratified sample, the performance that may reasonably be expected on another stratified or unstratified sample *drawn from the same distribution as the base data*. However, such alternative sampling from a single distribution is very different from the situation in which the class distribution changes between the distribution from which the base data are drawn to the distribution from which the target data are drawn.

The only other reason that we might expect $P(\pi_\lambda(X) \mid Y \wedge \cdot)$ to remain invariant across changing data distributions is if the values of $\pi_\lambda(X)$ are determined by the values of Y or the values of both $\pi_\lambda(X)$ and Y are determined by those of a third factor Z in such a manner that ensures the invariance. This is 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 $\pi_\lambda(X)$ will be causally affected by the presence or absence of fraud.

However, there are many other circumstances where it is not credible. For example, 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.

But even for the circumstance where there is a causal relationship from Y to X, $P(\pi_\lambda(X) \mid Y \wedge \cdot)$ may still vary from base to target data. 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 true 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 $P(\pi_\lambda(X) \mid Y \wedge \cdot)$ will change and ROC analysis can be expected to fail to accurately predict model performance.

As a final scenario, consider the circumstance where X contains both attributes A whose values are affected by those of Y, and attributes B whose values affect those of Y. For example, the disease, or fraud, might be more prevalent among a specific age group [$P(Y)$ is influenced by age] as well as causing specific symptoms or behaviors. 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, $P(\pi_\lambda(X) \mid Y \wedge \cdot)$ will change from the base to the target data and ROC analysis can be expected to fail to accurately predict model performance.

It might be argued that any change in $P(X \mid Y \wedge \cdot)$ constitutes concept drift and that it is not reasonable to expect any technique to accurately predict classification performance under concept drift. However, concept drift is defined conventionally as variation in a function from X to Y (for example, Bartlett, Ben-David, & Kulkarni, 2000). It follows that concept drift occurs when $P(Y \mid X \wedge \cdot)$ varies. Our example in Section 3 illustrates a concept that is defined as a function from X to Y. As we show, when $P(Y \mid \cdot)$ changes and the concept, $Play$ if and only if $Pleasant$ and $Free$, does not change, for $P(X \mid Y \wedge \cdot)$ to remain invariant requires a precise change in $P(X \mid \cdot)$. As we point out, for ROC analysis to provide accurate predictions of classification

performance under varying class distributions in this case requires that the world be organized such that the weather will only change in precise relationship to an individual's non-golfing commitments. As this illustrates, to maintain that ROC analysis should be expected by default to provide accurate predictions of classification performance under varying class distributions in the absence of concept drift (as conventionally defined) requires one to maintain that any change in $P(Y \mid \cdot)$ will be matched by a precise corresponding change in $P(X \mid \cdot)$ such that neither $P(X \mid Y \wedge \cdot)$ nor $P(Y \mid X \wedge \cdot)$ varies.

5. Conclusions

We have argued that there are conditions under which ROC analysis will not accurately predict model performance under varying class distributions. We have provided a detailed example that illustrates such conditions. If $P(X \mid Y \wedge \cdot)$ is invariant, any change in the class distribution $P(Y \mid \cdot)$ requires a change in at least one of $P(Y \mid X \wedge \cdot)$ or $P(X \mid \cdot)$. ROC analysis assumes that the likelihood function $P(\pi_\lambda(X) \mid Y \wedge \cdot)$ remains invariant. We have argued that unless the model has been formed from a stratified sample drawn from the same XY distribution as the target data it is not realistic to expect ROC analysis to accurately predict model performance under varying class distributions when either:

- there are causally relevant sub-categories of the Y values whose distributions vary at different rates between the base and target data, or

- attributes of X that are used by the model are causally related to Y such that the Y values depend on the X values.

A further constraint on the likely accuracy of ROC analysis under varying class distributions is the possibility that the variation might result from concept drift.

ROC techniques provide valuable means of assessing the potential trade-offs between true and false positive rates under varying model thresholds, and hence of a model's performance under varying misclassification cost functions. However, before practitioners use ROC techniques to compare the expected performance of alternative learning algorithms in the face of potential changes in class distribution, we believe that it is incumbent upon them to assure themselves that $P(\pi_\lambda(X) \mid Y \wedge \mathit{target}) \approx P(\pi_\lambda(X) \mid Y \wedge \mathit{base})$.

Acknowledgements

This paper has benefitted greatly from detailed feedback and suggestions by the anonymous reviewers. We are also very grateful to Tom Fawcett and Peter Flach for providing helpful and constructive comments. We are particularly grateful to the action editor, Foster Provost, who has provided very detailed and thoughtful comments and has done much to shape the paper's final form. Were he not our action editor, his contribution would merit acknowledgement as co-author.

References

Adams, N. M. and Hand, D.J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, *32*, 1139–1147.

Bartlett, P., Ben-David, S., & Kulkarni, S. (2000). Learning changing concepts by exploiting the structure of change. *Machine Learning*, *41*, 153–174.

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.

Flach, P.(2003). The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. *Proceedings of The Twentieth International Conference on Machine Learning.*

Provost, F. and Fawcett, T. (1997). Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 43–48.

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.

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.