
Machine Learning, 58, 33–38, 2005

A Response to Webb and Ting's *On the Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions*

TOM FAWCETT tom.fawcett@hp.com
HP Laboratories, 1501 Page Mill Road, Palo Alto, CA, USA

PETER A. FLACH Peter.Flach@bristol.ac.uk
Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK

Abstract. In an article in this issue, Webb and Ting criticize ROC analysis for its inability to handle certain changes in class distributions. They imply that the ability of ROC graphs to depict performance in the face of changing class distributions has been overstated. In this editorial response, we describe two general types of domains and argue that Webb and Ting's concerns apply primarily to only one of them. Furthermore, we show that there are interesting real-world domains of the second type, in which ROC analysis may be expected to hold in the face of changing class distributions.

Keywords: classification, classifier evaluation, ROC, class skew

1. Introduction

ROC graphs have become increasingly popular in machine learning, primarily because they seem to offer a flexible and robust framework for evaluating classifier performance (Fawcett, 2003; Flach, 2004). One of the main benefits of ROC graphs is their ability to separate error cost considerations from classifier performance. This ability has become important with the emergence of cost-sensitive learning as a popular research topic. A second benefit is the ability of ROC curves to remain invariant under changing class distributions; that is, as one class becomes more or less prevalent, classifier performance as depicted by an ROC graph should not change (Provost & Fawcett, 2001). This ability has become important for investigating domains in which class proportions change over time.

It is this second property that Webb and Ting (2005) challenge in an article in this issue. They present a simple domain and assert that only under very rare circumstances will ROC analysis continue to hold when class distributions are changed. In short, Webb and Ting challenge the assumption that class distributions may change without an underlying change in the class-conditional distributions (and, consequently, in the model's true and false positive rates) as well. They advise researchers to test this assumption on real world domains.

We are sympathetic to challenges to common assumptions in machine learning research (Provost, Fawcett, & Kohavi, 1998). The field could probably benefit from more questioning of its assumptions, and recommending that researchers test their assumptions in practice

is always good advice. However, we believe that Webb and Ting have overstated their criticism. In this article we revisit the case made by Webb and Ting and attempt to offer some additional insight into the issue of how changing class distributions may influence a model's performance.

2. Two main types of domains

In this section we describe two main types of classification domains, only one of which is scrutinized in Webb and Ting's analysis. They focus on a simple example domain, taken from Quinlan (1987), in which the goal is to determine when a golf enthusiast may play. There are two independent variables or observables, *Playing Conditions* (values *Pleasant* and *Unpleasant*) and *Prior Commitments* (values *Busy* and *Free*). The dependent variable *Golf* has values *Play* and *Don't Play*. The target concept is *Play* iff *Free* $\wedge$ *Pleasant*.

As a causal network, the "golf domain" may be depicted schematically as in figure 1(a), in which values of the observables $x_1, x_2, \ldots, x_n$ influence the class value Y. We call this an "$X \rightarrow Y$" domain.

Webb and Ting point out that in such $X \rightarrow Y$ domains, where the class is causally dependent upon X, it is unlikely that changes in the base rate of Y will result in the same true and false positive rates. This is true, and it is worth pointing out that in a $X \rightarrow Y$ domain this is unlikely. But Webb and Ting believe this observation implies that ROC analysis is generally suspect.

At least one other configuration is possible, shown in figure 1(b). In this configuration, the class value Y determines the probability of each observable x_i. This corresponds to a mixture modeling interpretation of changing priors: In a two-class classification problem, there are populations C1 and C2 comprising the sets from which examples of each class are drawn. The combined population is a mixture of samples drawn randomly from C1 and C2 at a given rate. We call such a domain a "$Y \rightarrow X$" domain. The classifier tries to distinguish C1 and C2 within the mixed population.

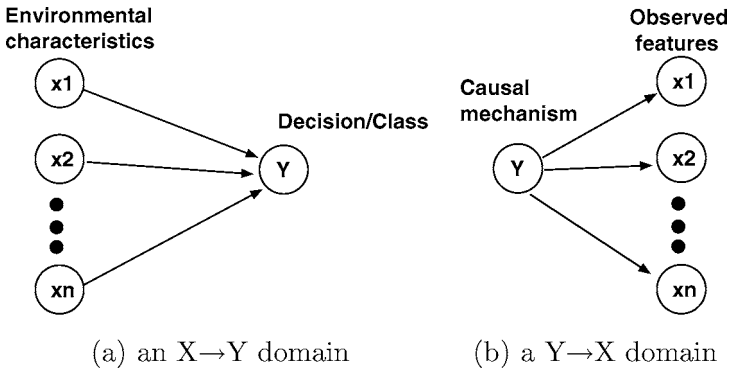

Figure 1. Schematics of two domain types.

In a $Y \rightarrow X$ domain, class priors change when C1 and C2 are sampled at rates different from their previous base rates. This is what Webb and Ting mean by "stratified sampling," though this phrase implies that the researcher alone is doing the sampling and is choosing to alter the class proportions. But stratified sampling may occur quite naturally as the phenomena producing instances in C1 and C2 become more or less active. In the next section we give a few examples of this.

3. Example domains

It is impossible to state whether $X \rightarrow Y$ or $Y \rightarrow X$ domains should be more prevalent in machine learning research. However, it is worth pointing out several realistic $Y \rightarrow X$ domains of interest.

Consider a medical diagnosis domain in which an infectious organism (Y) causes illness in humans. The task is to detect whether a given person is infected or not. Symptoms of infection (X) include fever, coughing and weakness. Each of these symptoms has some base level of prevalence in a population. As the organism spreads (for example, in an epidemic), the prevalence of infected patients ($Y = 1$) increases resulting in a shifting class distribution. The symptoms X will increase in the patient population as well. Barring a mutation of the organism, the symptoms should continue to occur in the same proportion relative to each other. An increase in the organism's prevalence should produce a proportional increase in its symptoms.

Consider an example of manufacturing fault detection. A manufacturing line produces fluorescent tubes by depositing a coating on the glass tubes with a heated gas vapor. As fluorescent tubes come off the line they are given a series of tests (X) to determine, among other things, whether the coating fluoresces properly. The vapor element, when under-heated ($Y = 1$), produces defective, uneven coatings with poor fluorescence, and such defects result in a set of off-normal test readings of the X values. As heating elements age they more frequently fail to heat properly. The base population contains a certain number of such defective instances. In the target population, elements have aged further and a higher number of defects are present, so the "defect" population is effectively sampled at a higher frequency. The appearance of the defect does not change. This is precisely a $Y \rightarrow X$ phenomenon.

It is important to note that we are not claiming that concept drift[1] cannot occur in these domains. Indeed, organisms mutate and equipment can degrade in new ways, resulting in concept drift. With these examples we are demonstrating that class distributions may easily change without being accompanied by the changes in FP and TP rates that Webb and Ting describe in their golf example. That is, the marginal class distributions $p(C)$ may change while the class-conditioned feature probabilities $p(X \mid C)$ remain the same. Such domains are exempt from the criticisms of Webb and Ting (2005), and the assumptions underlying ROC graphs continue to hold.

However, it is worth pointing out a final $Y \rightarrow X$ domain that illustrates a complexity. In cellular phone fraud detection (Fawcett & Provost, 1996; Fawcett & Provost, 1997), there is a population of legitimate users and a population of fraudsters. The task is to determine whether a legitimate account has been defrauded (Y). The population of account

behavior (X) is a mixture of the behavior of these two populations. The populations are generally unrelated and may vary over time due to, for example, promotions, competition from other carriers, or security vulnerabilities in the cellular system. Without the fraudulent or legitimate behavior fundamentally changing, we can observe variations in the amounts—and corresponding variations in probability distribution over the observed features.

But within each user population there may be sub-populations of users. For example, legitimate users may comprise business users, teens, emergency-only users, etc. Fraudsters may also have their own sub-populations. Each sub-population may have different behavior, resulting in different probability distributions over the observed variables. Formally, each class C_i may have a sub-population $C_{i,j}$ and $p(X \mid C_{i,j})$ may be different for different j's. An increase in a class prior $p(C_i)$ may be due to uneven increases across the sub-populations of class C_i. These changes may result in a varying $p(X \mid C_i)$. In such cases, the TP and FP rates may indeed change.

This case is what Webb and Ting refer to in their Section 4 as a "superclass of related sub-classes", though we prefer to call it a disjunctive target concept. When the disjunct prevalences change it is indeed problematic for ROC curves even though the domain is $Y \rightarrow X$.

4. Unknown class distributions

Finally, there is another advantage to using ROC curves with skewed class distributions that deserves specific mention. In some circumstances, the actual skew of the classes under which a classifier will be used may be simply *unknown* at learning time. In this case, the researcher must sample the P and N classes at some proportion in order to build up populations of each. The class-conditional probabilities $p(X \mid class)$ are unknown but fixed; in other words, the concept may be learned from the samples but it will not drift. The true class distribution may not be known until the classifier is deployed.

Under these circumstances, the researcher may perform stratified sampling from P and N, varying the class proportions and learning classifiers from the combined examples. Webb and Ting mention this in Section 4 of their article but it is worth recasting it in a positive light. ROC graphs are a useful tool for visualizing the performance of each classifier over the full range of class skews and cost assumptions. This is true whether the domain is $X \rightarrow Y$ or $Y \rightarrow X$.

5. Conclusions

Webb and Ting argue that an important assumption underlying the use of ROC graphs—that they remain invariant under changing class distributions—should never be assumed without reason. They imply that ROC proponents have overstated this claim. We argue, in response, that there are two important classes of domains and that Webb and Ting's concerns apply primarily to only one of them. Furthermore, we show that there are interesting real-world domains of the second type, in which ROC analysis may be expected to hold in the face of changing class distributions.

It is important to remember the context in which ROC analysis was promoted by machine learning researchers. Popular evaluation metrics such as classification accuracy and even precision-recall graphs were seen as inadequate for the challenges of some real-world domains. ROC graphs were recommended as a superior alternative (Provost & Fawcett, 1997). In spite of Webb and Ting's caveats, we believe they still are.

We are sympathetic to challenging assumptions underlying machine learning research, but when making such a challenge it is important to address three questions:

1. *How often may we expect this assumption to be violated in practice?*
2. *When the assumption is violated, what are the consequences?*
3. *If we should avoid making this assumption, what should we do instead?*

Questions one and two may be difficult to evaluate but they help the research community judge how serious the issue really is. Together they help researchers estimate the "expected penalty" of the violated assumption. Question three raises the issue of alternatives: unless a better alternative exists, researchers have no reasonable incentive to change their practice. We hope that this article contributes to a dialog in which these questions can be answered.

Acknowledgments

We thank Foster Provost for suggesting this editorial response. Discussions with Geoff Webb, Rob Holte and Foster Provost were very useful for clarifying some of the issues involved.

Note

1. We adopt the conventional definition of concept drift as a fundamental change in the relationship between independent variables and the dependent (class) variable. This is equivalent to the definition accepted by Webb and Ting (Webb and Ting, 2005).

References

Fawcett, T. (2003). ROC Graphs: Notes and practical considerations for researchers. Tech Report HPL-2003-4, HP Laboratories. Available: http://www.hpl.hp.com/personal/Tom_Fawcett/papers/ROC101.pdf.

Fawcett, T., & Provost, F. (1996). Combining data mining and machine learning for effective user profiling. In Simoudis, Han, & Fayyad (eds.). *Proceedings on the second international conference on knowledge discovery and data mining*. (pp. 8–13). CA AAAI Press: Menlo Park.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, *1:3*, 291–316.

Flach, P. (2004). The many faces of ROC analysis in machine learning. ICML-04 Tutorial. Notes available from http://www.cs.bris.ac.uk/∼flach/ICML04tutorial/index.html.

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 (KDD-97)* (pp. 43–48). Menlo Park, CA.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42:3*, 203–231.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In J. Shavlik (ed.), *Proceedings of ICML-98* (pp. 445–453). San Francisco, CA.

Quinlan, J. R. (1987). Learning decision trees. *Machine Learning*, *1:1*, 1–25.

Webb, G. I., & Ting, K. M. (2005). On the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning* (this volume).