An Analysis of Performance Measures For Binary Classifiers

Submitted for Blind Review

***Abstract*—If one is given two binary classifiers and a set of test data, it should be straightforward to determine which of the two classifiers is the superior. Recent work, however, has called into question many of the methods heretofore accepted as standard for this task. In this paper, we analyze seven ways of determining if one classifier is better than another, given the same test data. Five of these are long established and two are relative newcomers. We review and extend work showing that one of these methods is clearly inappropriate, and then conduct an empirical analysis with a large number of datasets to evaluate the real-world implications of our theoretical analysis. Both our empirical and theoretical results converge strongly towards one of the newer methods.**

***Keywords*-Classifier Evaluation; Performance Metrics; Supervised Learning;**

I. Introduction

Given two trained classifiers and a set of testing data, is it possible to determine which one is better? It would certainly seem so, but there has been much work in recent years that has exposed numerous potential difficulties within this problem. A natural measure to evaluate a classifier on a set of test data is *accuracy*; that is, the number of instances in the test set on which the classifier's prediction is correct divided by the total number of instances. While this seems acceptable as a performance measure, it is riddled with so many problems [1], [2], [3], [4] that we will not even bother to discuss it here.

In this paper, we examine four possible methods of determining if one classifier is superior to another given the same test data. The first three are the *F1-measure*, *average precision*, and *precision at k*, popular tools for performance measurement in the information retrival and natural language processing communities. The fourth is the *phi coefficient*, which has gained more traction in medical and psycological communities. The fifth is the *AUC*, a standard for many years that has come under attack recently [5], [6] regarding its validity. The sixth and seventh are the *AUK* and the *H-measure*, two performance measures that purport to solve some of the problems of the AUC. We examine the methods both mathematically and empirically, and find some to be more appealing than others on both counts.

The rest of this paper is organized as follows: Section II introduces notation and formally defines each of our performance measures. Section III reviews mathematics leading to the conclusion that the AUC as a performance measure is *incoherent*, then extends this mathematics to show that the same notion of incoherence applies to the AUK as well. Section IV describes some simple empirical tests that we perform to compare the results of our mathematics to results in the real world. Section V analyzes the outcome of these experiments and Section VI concludes.

II. Preliminaries

Consider that we have a set of training data, composed of a set of vectors of real values, each vector associated with a label $\in \{0, 1\}$, where we call 0 the "positive label" and 1 the "negative label". Now suppose we learn a classifier H on this training set that, given some test vector $\mathbf{x}$, outputs a single real value as its prediction, $H : \mathbf{x} \mapsto \Re$. Without loss of generality, we consider that when the output value, or "score" is lower, this indicates a higher probability of belonging to the positive class. Further suppose that we have a test set, $\mathcal{T}$, also composed of n labeled vectors, or test instances. Given this test set, we can run each instance through H and create a vector of scores, $\mathbf{s} = s_1, s_2, \ldots, s_n$, with each score $s_i \in \mathbf{s}$ corresponding to a classifier prediction on some instance within the test set.

In keeping with the notation of [5], we can view the scores of the positive instances in the test set as being drawn from a probability distribution $f_0(s)$, with associated cumulative distribution function $F_0(s)$, and similarly for the negative class, with associated functions $f_1(s)$ and $F_1(s)$. We also define the empirical probabilities within the test set of the positive and negative classes, respectively, as π_0 and π_1. Finally, we can define the overall distribution of scores, $f(s) = \pi_0 f_0(s) + \pi_1 f_1(s)$, and the associated cumulative distribution $F(s)$.

Now consider some threshold t on this distribution, such that instances with a score $s < t$ will be predicted to be positive by the classifier. With a given t, the familiar notion of *false positive rate* can be defined as $f_p(t) = F_1(t)$, and similarly the *true positive rate* or *recall* $r(t) = F_0(t)$. The *precision* is the number of true positive predictions divided by the total number of positive predictions, $p(t) = \pi_0 F_0(t)/F(t)$. The *accuracy* is proportion of correct predictions, $a(t) = \pi_0 F_0(t) + \pi_1(1 - F_1(t))$.

In what follows we will group our measures into two broad categories: The first, containing the F1-measure, the phi coefficient, and precision at k will fall into the category we will call *point measures*. That is, these measures give a performance estimate at a particular threshold, ignoring all others. The second category, containing all other measures, we will call *integrated measures*, measures which give a performance estimate by integrating over possible thresholds. We will examine the point measures first.

A. *The F1-Measure*

With precision and recall defined, it is an easy matter to define our first performance measure, the *F1-measure*[1] [7]. The F1-Measure is, given some value of t, the balanced harmonic mean of precision and recall:

$$M_{F1}(t) = \frac{2pr}{p+r} = \frac{\frac{2\pi_0 F_0^2(t)}{F(t)}}{F_0(t) + \frac{\pi_0 F_0(t)}{F(t)}}$$
$$= \frac{2\pi_0 F_0(t)}{F(t) + \pi_0}$$

The F1-measure is appealing in that the distribution of scores is not necessary for its calculation. We need only to know the number of true positives, false positives, and false negatives given by some classifier. In addition, it is more useful than accuracy as poor performance in terms of either precision or recall on the positive class leads to a low number.

If we have a distribution of scores, however, the F1-measure loses some of its appeal in that we need to pick a threshold to generate it. The most sensible choice, given our score distributions, would seem to be the one that maximizes the measure:

$$t_{F1} = \operatorname*{argmax}_t M_{F1}(t)$$

One way of comparing different classifiers, then, is to compare the values of $M_{F1}(t_{\max})$ on the score distributions generated by each classifier. Note that the values of $t_{\max}$ may be different for different classifiers, but in each case we choose a single $t = t_{\max}$, the one that optimizes performance, at which to evaluate each one.

B. *The Phi Coefficient*

The Phi Coefficient [8], also known as the *Matthews Correlation Coefficient* is another threshold-based measure of quality. Like the F1-Measure, it is designed to work on data where the class distribution is skewed. For a binary classifier, in our notation, the measure can be computed as follows:

$$\phi(t) = \frac{a(1-a)}{\sqrt{\pi_0 \pi_1 n^2 F(t)(1-F(t))}} \qquad (1)$$

Again, because we must select a threshold for the measure, it seems appropriate to select the threshold at which the measure is maximized:

$$t_\phi = \operatorname*{argmax}_t \phi(t)$$

and using $\phi(t_\phi)$ to compare classifiers. We leave a full discussion of ϕ to [8], but note anecdotally that while the measure has gained some traction in the biological and psychological communities, it is still not widely used in the machine learning community.

C. *r-Precision*

A collection of measures that we refer to collectively as *precision at k* measures are popular in information retrieval [9]. To calculate precision at k, we find t_k such that $nF(t_k) = k$, so exactly k examples are labeled positively. The precision at k is then $p(t_k)$.

The choice of k is somewhat arbitrary in the literature, with authors often reporting values for precision at several values of k. Often, the domain will dictate the appropriate value of k, such as search problems in which ten results per page may be presented. Another popular choice is precision at r or *r-precision* [10], where $r = n\pi_0$, the total number of positive examples.

D. *The Area Under the ROC Curve (AUC)*

While the F1-measure gives us a good way to measure performance, it forces us to make an explicit assumption about the relative costs of misclassifying examples of each class. The F1-measure assumes, by the definition of the balanced harmonic mean, that precision and recall are equally important. It is possible that this is not the case, and in cases where one is significantly more important than the other, performance may be dramatically different. We would like a measure that does not require a choice of threshold.

The area under the ROC curve fits this requirement [1]. An ROC curve is a plot of $F_0(t)$ vs. $F_1(t)$ for varying t. The ideal ROC curve is a right angle at $(0,1)$ so that there is a choice of threshold that perfectly separates positive and negative instances. An ROC curve that classifies the data randomly is a diagonal line from $(0,0)$ to $(1,1)$. A natural notion of performance is the area under this curve, which we will call the AUC. Following [5], we set $v = F_1(t)$ and define the *AUC* as:

$$\int_0^1 F_0(F_1^-1(v))dv$$

Noting that $dF_1(t)/dt = f_1(t)$, we can change the variable of integration to t, giving the AUC in terms of the score distributions:

$$\text{AUC} = \int_{-\infty}^{\infty} F_0(t) f_1(t) dt$$

We now compute performance by integrating over the range of possible thresholds. Effectively, this measure averages performance over all levels of specificity, and so we are relieved of the burden of having to choose a threshold.

An important drawback here is that all levels of specificity are often not equally likely in practice. In the case where negative examples far outnumber positive ones (as in many detection problems), high levels of specificity are typically required to produce a useful system. Yet, even the empirical

[1] Note that this F1 is not directly related the cumulative distribution function of the negative class $F_1(s)$. In order to remain consistent with [5] and avoid too much ambiguity, we use the non-traditional notation $M_{F1}(t)$ to refer to the F1-Measure at a given threshold t.

data distributions, defined by π_0 and π_1 are not used when computing the AUC. Some authors [11], [12] opt to resolve this difficulty by computing only a portion of the AUC, leading to the *a priori* choice of exactly which is the relevant portion.

On top of this, there is significant area under the ROC curve that is captured even by the random classifier, which seems intuitively incorrect to include in our calculations [6]. The Gini coefficient [13], a linear transformation of the AUC, fixes this, but does so without considering the underlying class distribution of the test data. We introduce in Sections II-E and II-F two measures that attempt to fix these difficulties.

E. Average Precision (AP)

As defined previously, a natural measure of performance given a threshold t is the precision, $p(t) = \pi_0 F_0(t)/F(t)$, which expresses how often the classifier is correct when it predicts the positive class. Because we do not know the relative importance of precision and recall in a given application, a sensible step is to simply average the precision at all possible levels of recall. With v as above:

$$\text{AP} = \int_0^1 p(F_1^{-}1(v))dv = \int_{-\infty}^{\infty} p(t)f_1(t)dt$$

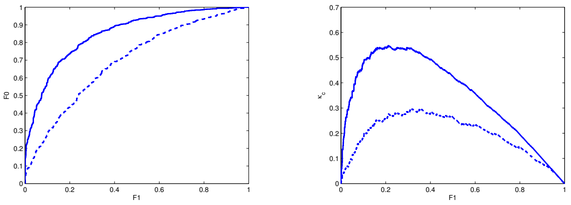
This measure, the average precision [10], [9], can also be interpreted geometrically as the area under a precision-recall curve. It is similar to the AUC in that it is an integration over all false positive rates, but different in that it is computed using π_0, and thus takes into account the class distribution of the test data.

F. The Area Under the Cohen's κ Curve (AUK)

Given some t, and our score distributions, we can also compute a measure known as *Cohen's κ* [14], or κ_c:

$$\kappa_c(t) = \frac{a(t) - p_c(t)}{1 - p_c(t)}$$

where $p_c(t) = \pi_0 F(t) + \pi_1(1 - F(t))$ is the probability of choosing the correct class by random chance, with the distribution of guesses determined by the selected t.



(a) Two ROC curves.

(b) The corresponding Cohen's κ curves.

Figure 1. A comparison of two ROC curves and their corresponding Cohen's κ curves.

It is important to note that, given the class distribution of the test data, κ_c is a function of t dictated completely by the distributions of scores. We acknowledge this by making it a function $\kappa_c(t)$. One can then plot a curve with $\kappa_c(t)$ on the vertical axis and $F_1(t)$ on the horizontal. This curve is differently shaped from the ROC (as seen in Figure 1), but a larger area underneath still indicates a superior classifier. We thus compute the area under this curve, the *AUK* [6], as a performance measure. With the same definition of v above:

$$\text{AUK} = \int_0^1 \kappa_c(F_1^{-}1(v))dv = \int_{-\infty}^{\infty} \kappa_c(t)f_1(t)dt$$

This measure still integrates over levels of specificity, treating them as equally likely, but uses κ_c to control for the distribution of the data. From the definition of κ_c, we can see that $\lim_{a \to p_c} \kappa_c(t) = 0$, so that classifiers with accuracy approaching random chance contribute increasingly less to the AUK. Like average precision, the AUK appears to take into account the class distribution of the test data while avoiding specific loss assumptions regarding false positives and false negatives.

G. The H-Measure

Another solution to the problems with the F-measure and the AUC is the *H-measure* [5]. Rather than integrating over levels of specificity, as do the AUC and AUK, the H-measure integrates over possible costs of misclassification. Suppose we have a function $Q(t; b, c)$ that computes the loss on the test set, where c (defined below) gives the relative cost of misclassification for positive and negative examples, b is a scaling factor, t is a threshold. If we have a distribution $u(c)$ over relative costs, and define $T(c)$ as in Equation 3 below, we can integrate over this distribution to compute an expected loss L_u:

$$L_u = \int Q(T(c); b, c)u(c)dc$$

By dividing by the maximum loss and subtracting from one, we have a measure on the same scale as the others reviewed. However, we have introduced the need to specify not just relative misclassification costs, but a *distribution* over costs. It is suggested in [5] that a symmetric beta distribution is a sensible assumption. But is this assumption necessary? After all, the AUC, the AUK, and average precision appear to have given us measures requiring no such assumptions. We will see in the following section, however, that this is not strictly true.

III. The Incoherence of Certain Performance Measures

Here we will first briefly review the mathematics of [5] that shows the sense in which the AUC is incoherent, which is done in Sections III-A and III-B. We then generalize this to show that this incoherence applies to the AUK and many related functions. We will pass over several of the subtle points in [5] in the interest of brevity, but encourage the reader to consult the original.

To conduct this examination, we introduce the variables c_0 and c_1, representing the cost of misclassifying a positive and a negative example, respectively. We also introduce $c = c_0/(c_0 + c_1)$ as a variable which expresses the *relative cost* of misclassification, and $b = c_0 + c_1$ to express the *scale* of the combined costs.

Crucially, we note that the true values for c_1 and c_0 are most certainly *properties of the classification problem*. For example, in medicine, the cost of missing a positive case, thereby allowing illness to go undetected, is often greater than giving a false positive diagnosis. In this case, $c_0 > c_1$. In contrast, algorithms for visual object detection may prefer to miss a single positive image rather than present a user with an image that does not contain the object, leading to $c_1 > c_0$. Regardless of what these costs are, they are dictated by the problem domain, and certainly not by the classification algorithm.

A. The Relationship Between Cost and Threshold

With costs c_0 and c_1 in hand, it is easy to compute the misclassification loss given some threshold t:

$$c_0\pi_0(1 - F_0(t)) + c_1\pi_1F_1(t) \tag{2}$$

Of course, given the ratio of the misclassification costs, there will be some $t = T(c)$ that minimizes the loss. Using c instead of c_0 and c_1:

$$T(c) = \operatorname*{argmin}_t c\pi_0(1 - F_0(t)) + (1 - c)\pi_1F_1(t) \tag{3}$$

The loss function Q for any t can be written as:

$$Q(t; b, c) = \{c\pi_0(1 - F_0(t)) + (1 - c)\pi_1F_1(t)\}b \tag{4}$$

Differentiating this expression and setting to zero gives an equation that can be solved for the minimizing value of t:

$$c\pi_0f_0(t) = (1 - c)\pi_1f_1(t)$$

with $d^2Q/dt^2 = 0$. There are a few problems with this process. First is that there may be many values of t that satisfy this expression. Second is that, in practice, F_0 and F_1 are discrete and therefore not differentiable. Both of these objections are dealt with in Section 5 of [5]. We assume for the rest of this analysis that each c can be matched one-to-one to a corresponding threshold t.

Given this one-to-one relationship, we can see that, given some threshold, one can easily solve for the cost $c(t)$ that is minimized under this threshold:

$$c(t) = \frac{\pi_1f_1(t)}{\pi_0f_0(t) + \pi_1f_1(t)} \tag{5}$$

We see then, that there is a correspondence between cost and threshold. Semantically, this shows that our belief about which threshold is appropriate implies a belief about the relative cost of misclassification.

B. The Incoherence of the AUC

Given Equation 4, which gives the loss given some threshold and misclassification costs, it is easy to calculate the expected loss even when the misclassification costs are unknown. All we require is a distribution $w(c)$ over relative misclassification costs. We then integrate over this distribution to get the expected loss:

$$L = \int_0^1 \{c\pi_0(1 - F_0(T(c))) + (1 - c)\pi_1F_1(T(c))\}w(c)dc$$

Of course, according to Section III-A, this is equivalent to integrating over a distribution over thresholds, $W(t)$:

$$L = \int_{-\infty}^{\infty} \{c(t)\pi_0(1 - F_0(t)) + (1 - c(t))\pi_1F_1(t)\}W(t)dt \tag{6}$$

where $c(t)$ is the cost implied by the threshold, as given in Equation 5. Again, the form of $w(c)$, and by extension $W(t)$, should certainly be dictated by the domain. However, consider the case where we define a weighting $W_G(t)$ that is instead dictated by the *distribution of scores given by the classifier*:

$$W_G(t) = \pi_0f_0(t) + \pi_1f_1(t) \tag{7}$$

Substituting both Equations 7 and 5 into Equation 6 yields the following expression for the weighted loss, L_G:

$$L_G = \int_{-\infty}^{\infty} \pi_0\pi_1\{f_1(t)(1 - F_0(t)) + f_0(t)F_1(t)\}dt$$

With a little bit of deft calculus in [5], we arrive at:

$$\begin{aligned} L_G &= 2\pi_0\pi_1 \int_{-\infty}^{\infty} \int_T^{\infty} f_1(t)f_0(s)dsdt \\ &= 2\pi_0\pi_1 \left(1 - \int_{-\infty}^{\infty} \int_{-\infty}^{T} f_0(s)dsf_1(t)dt\right) \\ &= 2\pi_0\pi_1(1 - \text{AUC}) \end{aligned}$$

We thus obtain a linear transformation of the AUC simply by computing the expected loss with different misclassification costs weighted according Equation 7. But Equation 7 defines a weighting based not on the problem domain, but the distributions f_0 and f_1 of the scores given by the classifier! The problem here is obvious: Our tool for measurement (the AUC) is different *depending on what is being measured.* This is the sense in which the AUC is incoherent. It is as if the meter were differently sized depending on whether wood or steel were being measured, and using measurements in such meters to compare the length of a wooden object with that of a steel one!

C. The Incoherence of Integrated Performance Measures

We have, in Sections III-A and III-B, reviewed the mathematics in [5] that shows the incoherence of the AUC as a performance measure. We show here, with a slight generalization of this argument, that the AUK and average precision are similarly incoherent.

To see this more clearly we first recall the weighting function $W_G(t)$, which was a function of the threshold t, from Equation 7. To this we add, for notational convenience, another function of the threshold, $W_I(t)$:

$$W_I(t) = f_1(t) - f_1(t)F_0(t) + f_0(t)F_1(t) \tag{8}$$

With this notation defined, we can prove the following:

Theorem 1. *For any performance measure M of the form:*

$$M = \int_0^1 m(F_1^{-}1(v))dv$$

Where $m(t)$ is a differentiable function of the threshold t, M is equivalent, within a multiplicative constant, to the expected loss with a distribution over thresholds (and by Equation 5, misclassification costs) given by:

$$W_m(t) = \frac{m(t)f_1(t)W_G(t)}{W_I(t)} \tag{9}$$

where $W_I(t)$ and $W_G(t)$ are given by Equations 7 and 8, respectively.

Proof: First, we change the variable of integration as we have done for each of the integrated measures:

$$M = \int_{-\infty}^{\infty} m(t)f_1(t)dt$$

Next, by substitution of Equation 7 into Equation 5:

$$c(t) = \frac{\pi_1 f_1(t)}{W_G(t)}$$

and by substituting this in turn into Equation 6:

$$\begin{aligned} L = \int_{-\infty}^{\infty} &\left\{ \frac{\pi_1 f_1(t)}{W_G(t)} \pi_0 (1 - F_0(t)) + \right. \\ &\qquad \left. \left(1 - \frac{\pi_1 f_1(t)}{W_G(t)}\right) \pi_1 F_1(t) \right\} W(t)dt \\ = \int_{-\infty}^{\infty} &\left(\frac{\pi_0 \pi_1 (f_1(t) - f_1(t)F_0(t) + f_0(t)F_1(t))}{W_G(t)} \right) W(t)dt \\ = \int_{-\infty}^{\infty} &\left(\frac{\pi_0 \pi_1 W_I(t)}{W_G(t)} \right) W(t)dt \end{aligned}$$

Substituting our constructed weighting $W_m(t)$ from Equation 9 for $W(t)$ gives the weighted loss L_m, and the prescribed conclusion:

$$\begin{aligned} L_m &= \int_{-\infty}^{\infty} \left(\frac{\pi_0 \pi_1 W_I(t)}{W_G(t)} \right) \left(\frac{m(t)f_1(t)W_G(t)}{W_I(t)} \right) dt \\ &= \pi_0 \pi_1 \int_{-\infty}^{\infty} m(t)f_1(t)dt \\ &= \pi_0 \pi_1 \mathrm{M} \end{aligned}$$

■

D. A Note About The Point Measures

We saw in Section III-C that measures of loss based integrating some quantity over all possible thresholds implies contradictory assumptions about the relative misclassification cost. A crucial step in this line of inference is that choosing a threshold value implies a belief about loss. What about the F1-measure, the phi coefficient, and precision at k?

Consider that each of these measures defines a loss[2]. Furthermore, though we omit the proof, it should be clear that each of these losses is differentiable with respect to t, as they are fairly simple algebraic functions of $F_0(t)$ and $F_1(t)$, which are themselves differentiable. We can thus use differentiation to solve for the thresholds that minimize each of these measures. Even more clearly, because our distribution of thresholds is in practice discrete and finite, we can simply test all possible thresholds to find the minimum loss.

In our use of the F1-measure and the phi coefficient, we compare the values of these measures *taken at the threshold at which they are maximized*. Thus, our assumptions about threshold and loss remain consistent. Note that this generalizes to the case where the thresholds are set based on cross validation or holdout data: If these measures are used for evaluation, it would be inexplicable to choose the thresholds that do not maximize these measures on the holdout data.

In contrast, precision at k selects an arbitrary threshold and allows this threshold to dictate the parameters of a strange loss function of the form of Equation 2 where $c_1 = 1/k$ and $c_0 = 0$. Note that $c = 0$ under the above definition of c, and thus this loss is trivially maximized at any t such that $F(t) = 0$; when no examples are predicted to be in the positive class. Clearly, this solution is not satisfactory, and there is thus some level of incoherance here: The threshold minimizing the loss appears to be incorrect.

Does this mean that precision at k is completely without use? No. The measure is valid in cases where the constant k is one that is *imposed by the domain*. A good example of this occurs in search, when one may return, say, 10 results per page. In this situation, precision at 10 is valid, as the domain is imposing a loss function of this form. Measures such as r-precision, in which r is set by the test data distribution, are not domain-imposed losses, but losses chosen by the evaluator.

E. Discussion

We see then, that the AUK and average precision suffer from the same notion of incoherence as the AUC: That it is equivalent to computing an expected loss, with the weighting of various misclassification costs controlled by

[2] For these measures, where the upper bound is unity and larger values mean better classifiers, the implied loss would be $1 - m(t)$, where m is the measure.

Dataset	**Features**	**Labels**	**Instances**
Landsat	36	6	6435
FBIS	2000	17	2459
Waveform	40	3	5000
Yeast	103	14	2417
Sound	52	10	3715

Table I
PROPERTIES OF THE DATASETS USED IN THESE TESTS.

the distribution of scores. In Section II, we saw that the integrated measures are specifically constructed to avoid having to make assumptions about the relative costs of misclassification. In fact, however, Section III shows that this is not the case at all. Not only have assumptions been made by these measures, but the assumptions are different for different classifiers operating on the same test data. While the measures are all significantly different, none fix this most crucial problem.

In contrast, the phi coefficient, F1-, and H- measures *do* make explicit assumptions about the misclassification cost. While this requires the individual researcher to make a choice, thus costing the measure some objectivity, it is clearly better than the inappropriate choice implicitly made by average precision, the AUC, and the AUK. In light of this, the main objection raised against the H-measure in [6] seems to be baseless. Other objections raised in the same work, that it is difficult to compute and has somewhat unclear semantics, might be considered a small price to pay for coherence.

IV. EXPERIMENTAL SETUP

In Section III, we reviewed and extended mathematical objections to some performance measures, but how often do these objections make a difference in practice? There has been previous work [15] suggesting that, at least, ROC curves of different classifiers on the same data rarely dominate each other entirely, which opens up the possibility of relative misclassification costs being at least somewhat important.

All of the measures in Section II purport to do the same thing: Given two classifiers, it will always assign a higher value to the one that is "better". As all of the notions of "better" espoused by these measures seem intuitively reasonable, one objective way to compare the measures is to see how often one disagrees with all others. As such, we have composed a collections of datasets, and have trained a collection of classifiers over these datasets. For each pair of classifiers on each dataset, we can compare them using each of the four measures, then examine the cases in which the measures disagree to assess the effectiveness of each measure.

The datasets are described below and some of the properties are given in Table I. The proportion of the positive class ranges from 0.9% to 73.5% with a median proportion of 11.0%

- The "Landsat Satellite" Statlog dataset used in [16] and several other papers, featuring multi-spectral images of the earth taken from a satellite. The classification task is to predict the type of vegetation or soil from the spectral values of a 3x3 patch of the image.
- The "FBIS" text data set from TREC-09 used in [17] and elsewhere, where the classification task is to predict the topic of a document given word counts.
- The waveform database generator data used in [18] and elsewhere. This is a synthetic dataset described in detail in [19].
- The multi-label yeast gene function dataset described in [20], which has become a benchmark in the multi-label learning literature.
- The multi-label sound classification dataset described in [21], in which small windows of sound are to be labeled with one or more of 10 semantic concepts, such as speech, music, crowd noise, etc.

On each of these datasets, we apply five different methods of dimensionality reduction in order to further vary the geometry of our data. In every case except k-means, the number of dimensions is determined by using PCA and selecting the smallest number of components such that 95% of the variance is captured. We use freely available implementations[3] of the following well-established methods, surveyed in [22]:

- Principle Components Analysis
- Kernel Principle Components Analysis
- Locally Linear Embedding
- Neighborhood Preserving Embedding
- k-means: Clustering the data into 32 means and using the distance from the point to each of the cluster centroids as the feature vector for that point.

This leads to six different representations of each data set, including the original unreduced representation. For each representation of each dataset, we treat each label as a one-vs.-rest binary labeling problem. On each problem, we train nine different classifiers: Naïve Bayes, logistic regression, SVMs with linear kernels, SVMs with RBF kernels, 1-nearest neighbor, 10-nearest neighbor, J48 decision trees, adaboosted J48 decision trees, and adaboosted decision stumps. For all of these, we use the default options in the Weka [23] data mining package. We then do pairwise comparisons of each pair of classifiers trained on a given problem, resulting in 36 comparisons per problem.

V. RESULTS

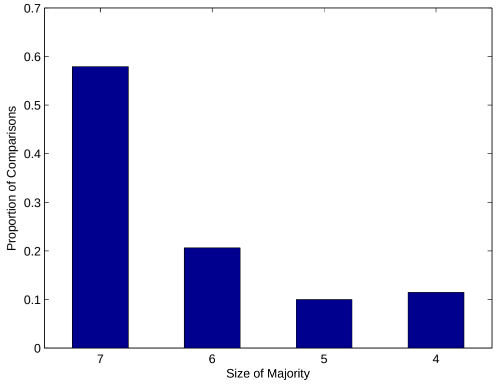
In Figure 2 we see the 11188 classifier comparisons broken out by level of agreement among the various performance measures. We see that, in about 58% of these com-

[3] http://cseweb.ucsd.edu/~lvdmaaten/dr/download.php

parisons, all measures agree on which of the two classifiers in the comparison is better. We would expect that, given that the goals of each measure are the same, that this would be the case.

On the negative side, in the other 42% of cases, we see that there is some level of disagreement among the measures, with either one, two, or three measures disagreeing with the majority. This means that, in nearly half of the comparisons we make, we can decide which of the classifiers is better simply by selecting a certain performance measure. This examination, then, seems to be of practical importance.

Figure 2. A breakdown of the 11188 classifier comparisons by level of agreement.



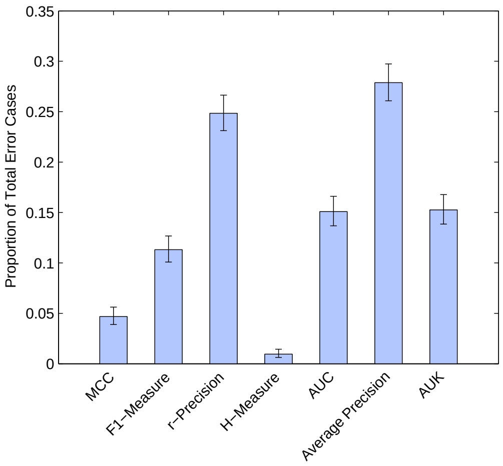
We now concentrate our examination on the roughly 2300 comparisons in which one of the measures disagrees with all others. These we will refer to as "error cases". Given that all of these measures have fairly deep precident in the literature, the assumption made here is that if the evaluation of one disagrees with that of the other six, this is because the other six measures have all detected something that the one has not. In this way, we can use number of error cases as way to quantify the "generality" of each measure.

The obvious null hypothesis here is that all measures are equally general, and that the assumptions, implicit or explicit, made by one will lead to no more error cases than the assumptions made by the others. With no other information, this seems reasonable.

Figure 3 shows us, unsurprisingly, that this is not the case. We use the Wilson score interval [24] to plot 95% confidence intervals on each bar, which shows us that the difference between the number of error cases for the various measures is indeed signifcant. We also use the G-test [25] to assure ourseleves that the difference between the empirical results and the null hypothesis is statistically significant, which it is ($G = 1095$, $p < 10^{-200}$).

The plot shows that the measures we identified as incoherant make up a larger proportion of error cases than the other measures. Importantly, the H-measure, engineered to make the most sensible assumption about loss, performs much better than all others, almost never disagreeing with all other measures. The phi coefficent (labeled MCC for Matthews Correlation Coefficient) also does well, dispite measuring performance at only a single threshold.

Figure 3. A comparison of performance measures by percentage of disagreements with all other measures.

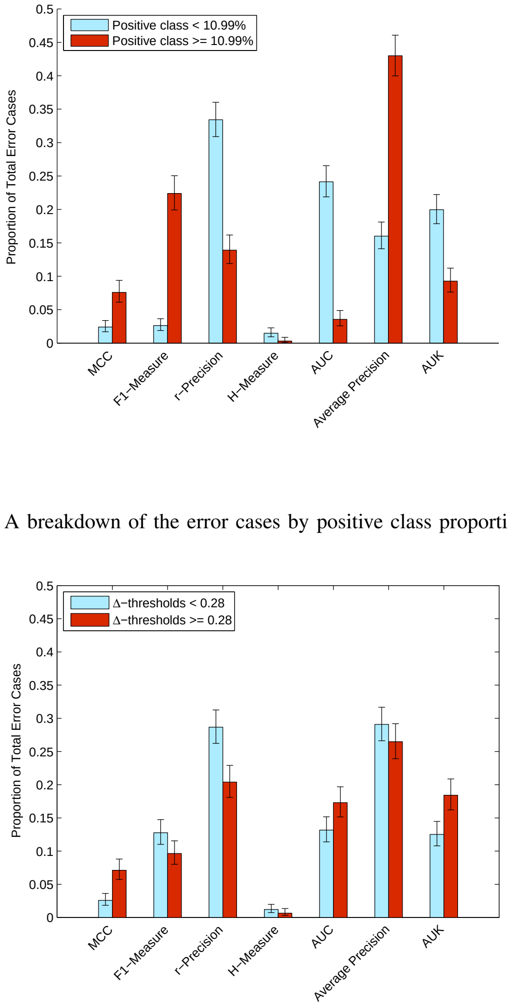


It is interesting to break down the error cases further in two ways. The first way, shown in Figure 4(a) is by the data distribution. We plot the cases where the proportion of the positive class in the test data is greater than the median proportion versus cases where it is less. As we see in the plot, the performance measures are extremely sensitive to the data distribution. The AUC, for example, makes up a far greater percentage of the errors when the data distribution is highly skewed than when it is balanced. The F1-measure, on the other hand, deals with skewed data much better than it does with balanced data. Only the H-measure has extremely low error rates in both cases.

The second breakdown, in Figure 4(b) is by the number of unique thresholds in the data. Because our score distributions are discrete, one way of measuring differences in the distributions is to count the number of unique thresholds in the data. If scores are duplicated, this number will be lower. Specifically, we measure the *difference* in the number of unique thresholds between the two score distributions being compared, normalized by the number of possible thresholds (that is, the total number of scores).

We see in Figure 4(b) that, while not as important as data distribution, this still seems to be a differentiating factor for some measures. In particular, the AUK and AUC seem to be sensitive to differences in the number of thresholds.

Finally, we would like to know if certain classifiers are more likely to give rise to error cases than others. We thus plot the total number of error cases which involve a given classifier (that is, in comparisons where either one classifier



(a) A breakdown of the error cases by positive class proportion

(b) A breakdown of the error cases by the difference in the number of unique thresholds of the score distributions of the two classifiers being compared.

Figure 4. Two different breakdowns of the error cases.

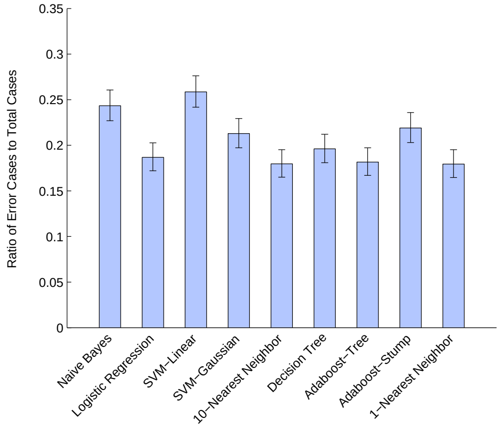
or the other is of the given type), normalized by the number of total comparisons involving that classifier.

We see in Figure 5 that there is significant overlap between the confidence intervals of many of the classifiers, indicating the type of classifier is not particularly important in identifying error cases. Crucially, this indicates that the choice of performance measure is an important issue regardless of the classifiers under comparison.

VI. Conclusions

It appears, then, we have some level of consistency between our theoretical and our empirical analysis. The measures that we identified as making assumptions about loss that are in some sense incoherant are shown, by our empirical analysis, to evaluate classifiers in a way that disagrees more often with other established measures. These "error cases" for each performance measure are dependant on the distribution of the test data and to a lesser extent on the number of thresholds in the distribution of scores given by the classifier. Finally, these error cases do not appear to be endemic to a particular classifier and thus all classifier comparisons are potentially suspect.

Figure 5. A comparison of the ratio of total comparisons involving a given classifier to comparisons involving that classifier where one performance measure disagrees with all others.



One implication of this work is that a researcher could engineer "better performance" by selecting the right performance measures. Given our analysis, this should be obvious. We have shown that all of these measures make some assumption about loss. By tuning the definition of loss, one can make a classifier appear superior. It is thus extremely important to note that, to our knowledge, none of these measures are "loss neutral" in any mathematical sense.

We have also shown that, paradoxically, using a coherant point measure such as F1 or ϕ is sometimes more general, in that it disagrees less often with all other measures, than ones that integrate over all thresholds. This points to the coherance of the loss assumption as a crucial component of any performance measure.

What of other performance measures? Some, like the popular *coverage* metric [26], appear to fall in the group of coherant point measures, in that the threshold implies a loss that is minimized at that threshold. Others, like the *11-point average precision* used in TREC competitions [27] appear to fall in the same category as the H-measure, specifying an explicit distribution over thresholds, then integrating over this distribution. We suspect that other measures may be shown to be making incoherant assumptions.

Finally, this work addresses a deep mathematical question about classifier evaluation: Is it possible to evaluate a classifier in a strictly loss-neutral manner? The answer, as far as we know currently, is an emphatic no. The alternative, then is to make an assumption about the loss that is as sensible as possible. In [5], we have not only the basis for these analyses, but also a well-formulated assumption in the proposed H-measure, the generality of which is borne out in our experiments. We recommend the H-measure for all future assessments of binary classification algorithms. As an alternative, the maximizing value of ϕ would also seem to be a coherant and sufficiently general choice.

References

[1] C. X. Ling, J. Huang, and H. Zhang, “AUC: a statistically consistent and more discriminating measure than accuracy,” in *IJCAI*, 2003, pp. 519–526.

[2] A. Ben-David, “About the relationship between ROC curves and Cohen’s kappa,” *Engineering Applications of Artificial Intelligence*, vol. 21, pp. 874–882, September 2008.

[3] F. J. Provost and T. Fawcett, “Robust classification for imprecise environments,” *Machine Learning*, vol. 42, no. 3, pp. 203–231, 2001.

[4] J. Demsar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1–30, January 2006.

[5] D. J. Hand, “Measuring classifier performance: a coherent alternative to the area under the ROC curve,” *Machine Learning*, vol. 77, pp. 103–123, October 2009.

[6] U. Kaymak, A. Ben-David, and R. Potharst, “AUK: a simple alternative to the AUC,” Erasmus Research Institute of Management (ERIM), Research Paper ERS-2010-024-LIS, Jun. 2010. [Online]. Available: http://ideas.repec.org/p/dgr/eureri/1765019678.html

[7] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[8] H. Cramer, *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[9] E. Yilmaz and J. A. Aslam, “Estimating average precision when judgments are incomplete,” *Knowledge and Information Systems*, vol. 16, pp. 173–211, July 2008.

[10] J. A. Aslam and E. Yilmaz, “A geometric interpretation and analysis of r-precision,” in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 664–671.

[11] D. McClish, “Analyzing a portion of the ROC curve,” *Medical Decision Making*, vol. 9, no. 3, pp. 190–195, 1989.

[12] L. E. Dodd and M. S. Pepe, “Partial AUC estimation and regression,” *Biometrics*, vol. 59, pp. 614–623, 2003.

[13] D. J. Hand and R. J. Till, “A simple generalisation of the area under the ROC curve for multiple class classification problems,” *Machine Learning*, vol. 45, pp. 171–186, October 2001.

[14] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, April 1960.

[15] F. Provost, T. Fawcett, and R. Kohavi, “The case against accuracy estimation for comparing induction algorithms,” in *In Proceedings of the Fifteenth International Conference on Machine Learning*, 1997, pp. 445–453.

[16] Z. Ghahramani and H.-C. Kim, “Bayesian classifier combination,” *Biomedical and Environmental Sensing*, vol. 38, no. 1, pp. 279–294, 2003.

[17] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, March 2003.

[18] G. Valentini and T. G. Dietterich, “Low bias bagged support vector machines,” in *International Conference on Machine Learning*. Morgan Kaufmann, 2003, pp. 752–759.

[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[20] A. Elisseeff and J. Weston, “A kernel method for multi-labelled classification,” in *Annual ACM Conference on Research and Development in Information Retrieval*, 2005, pp. 274–281.

[21] C. Parker, “An empirical study of feature extraction methods for audio classification,” in *ICPR ’10: The Twentieth International Conference on Pattern Recognition*, Istanbul, Turkey, August 2010.

[22] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, “Dimensionality reduction: A comparative review,” Tilburg University, Tech. Rep. TiCC-TR 2009-005, 2009.

[23] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*, 2nd ed., ser. Morgan Kaufmann series in data management systems. Morgan Kaufmann, 2005.

[24] E. B. Wilson, “Probable inference, the law of succession, and statistical inference,” *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.

[25] J. H. McDonald, *Handbook of Biological Statistics*, 2nd ed. Sparky House Publishing, 2009.

[26] S. Ji, L. Yuan, Y.-X. Li, Z.-H. Zhou, S. Kumar, and J. Ye, “Drosophila gene expression pattern annotation using sparse features and term-term interactions,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 407–416.

[27] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.