An Analysis of Performance Measures For Binary Classifiers

Charles Parker
BigML, Inc.
2851 NW 9th St.
Corvallis, OR 97330
parker@bigml.com

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 seven 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 retrieval and natural language processing communities. The fourth is the phi coefficient, which has gained more traction in medical and psychological 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 these 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 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 \( x \), outputs a single real value as its prediction, \( H : x \rightarrow \mathbb{R} \). 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, \( T \), 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, \( s = s_1, s_2, \ldots, s_n \), with each score \( s_i \in s \) 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 \( \pi_0 \) and \( \pi_1 \).

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 \( F(s) \).
\[ \pi_0 F_0(t) / F(t) \]. 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:

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

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 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:

\[ t_{F1} = \text{argmax}_t M_{F1}(t) \]

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

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

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

\[ t_\phi = \text{argmax}_t \phi(t) \]

and using \( \phi(t_\phi) \) to compare classifiers. We leave a full discussion of \( \phi \) 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 data distributions, defined by \(\pi_0\) and \(\pi_1\) 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 \textit{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 as it integrates over all false positive rates, but different in that it is computed using \(\pi_0\), and thus takes into account the class distribution of the test data.

F. The Area Under the Cohen’s \(\kappa\) Curve (AUK)

Given some \(t\), and our score distributions, we can also compute a measure known as Cohen’s \(\kappa\) [14], or \(\kappa_c\):
\[
\kappa_c(t) = \frac{a(t) - p_c(t)}{1 - p_c(t)}
\]
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 \(t\).

It is important to note that, given the class distribution of the test data, \(\kappa_c\) 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 \(\kappa_c\) to control for the distribution of the data. From the definition of \(\kappa_c\), we can see that \(\lim_{u \to p_c} \kappa_c(t) = 0\), 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 \textit{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 \textit{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 negative example, respectively. We also introduce $c_0$ and $c_1$ to the domain, and certainly not by the classifier 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 $t$:

$$c_0 \pi_0 (1 - F_0(t)) + c_1 \pi_1 F_1(t)$$

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 $c_1$:

$$T(c) = \arg\min_t c \pi_0 (1 - F_0(t)) + (1 - c) \pi_1 F_1(t)$$

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

$$Q(t; b, c) = \left\{ c \pi_0 (1 - F_0(t)) + (1 - c) \pi_1 F_1(t) \right\} b$$

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

$$c \pi_0 f_0(t) = (1 - c) \pi_1 f_1(t)$$

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:

$$c(t) = \frac{\pi_1 f_1(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)}$$

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:

$$L = \int_0^1 \left\{ c \pi_0 (1 - F_0(T(c))) + (1 - c) \pi_1 F_1(T(c)) \right\} w(c) dc$$

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

$$L = \int_{-\infty}^{\infty} \left\{ c(t) \pi_0 (1 - F_0(t)) + (1 - c(t)) \pi_1 F_1(t) \right\} W(t) dt$$

(7)

where $c(t)$ is the cost implied by the threshold, as given in Equation 6. 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:

$$W_G(t) = \pi_0 f_0(t) + \pi_1 f_1(t)$$

(8)

Substituting both Equations 8 and 6 into Equation 7 yields the following expression for the weighted loss, $L_G$:

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

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

$$L_G = 2 \pi_0 \pi_1 \int_{-\infty}^{\infty} \int_0^t f_1(t) f_0(s) ds dt$$

$$= 2 \pi_0 \pi_1 \left[ 1 - \int_{-\infty}^{\infty} \int_0^t f_0(s) f_1(t) ds dt \right]$$

$$= 2 \pi_0 \pi_1 (1 - \text{AUC})$$

We thus obtain a linear transformation of the AUC simply by computing the expected loss with different misclassification
costs weighted according Equation 8. But Equation 8 defines a weighting based on 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.

Here we come to the crucial point: Given two different classifiers trained and tested on exactly the same data, the AUC is actually measuring a different quantity for each classifier. 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 8. To this we add, for notational convenience, another function of the threshold, \( W_1(t) \):

\[
W_1(t) = f_1(t) - f_1(t)F_0(t) + f_0(t)F_1(t)
\]

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 6, misclassification costs) given by:

\[
W_m(t) = \frac{m(t)f_1(t)W_G(t)}{W_1(t)}
\]

where \( W_1(t) \) and \( W_G(t) \) are given by Equations 8 and 9, 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 8 into Equation 6:

\[
c(t) = \frac{\pi_1f_1(t)}{W_G(t)}
\]

and by substituting this in turn into Equation 7:

\[
L = \int_{-\infty}^\infty \left( \frac{\pi_1f_1(t)}{W_G(t)} \right) \pi_0(1 - F_0(t)) + (1 - \frac{\pi_1f_1(t)}{W_G(t)}) \pi_1F_1(t) W(t)dt
\]

\[
= \int_{-\infty}^\infty \left( \frac{\pi_0\pi_1(f_1(t) - f_1(t)F_0(t) + f_0(t)F_1(t))}{W_G(t)} \right) W(t)dt
\]

\[
= \int_{-\infty}^\infty \left( \frac{\pi_0\pi_1W_1(t)}{W_G(t)} \right) W(t)dt
\]

Substituting our constructed weighting \( W_m(t) \) from Equation 10 for \( W(t) \) gives the weighted loss \( L_m \), and the prescribed conclusion:

\[
L_m = \int_{-\infty}^\infty \left( \frac{\pi_0\pi_1W_1(t)}{W_G(t)} \right) \left( \frac{m(t)f_1(t)W_G(t)}{W_1(t)} \right) dt
\]

\[
= \pi_0\pi_1 \int_{-\infty}^\infty m(t)f_1(t)dt
\]

\[
= \pi_0\pi_1 M
\]

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.

\(^2\)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.
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 incoherence 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 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, such as speech, music, crowd noise, etc. may 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Features</th>
<th>Labels</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat</td>
<td>36</td>
<td>6</td>
<td>6435</td>
</tr>
<tr>
<td>FBIS</td>
<td>2000</td>
<td>17</td>
<td>2459</td>
</tr>
<tr>
<td>Waveform</td>
<td>40</td>
<td>3</td>
<td>5000</td>
</tr>
<tr>
<td>Yeast</td>
<td>103</td>
<td>14</td>
<td>2417</td>
</tr>
<tr>
<td>Sound</td>
<td>52</td>
<td>10</td>
<td>3715</td>
</tr>
</tbody>
</table>

Table I

Properties of the datasets used in these tests.
• 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: Naive 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. In about 58% of these comparisons, all measures agree on which of the two classifiers in the comparison is better. We would expect, 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 precedent 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 a way to quantify the “generality” of each measure.

It is important to understand that the true value for c, or even the distribution \( u(c) \) is unknown for all of the test data. If one knew the correct choice for c or \( u(c) \), one could use that choice to evaluate classifiers; the measures discussed here all substitute assumptions for this choice. Said another way, it is only in the case where the true loss is unknown that the measures discussed here are applicable. Thus, we do not claim that a measure is wrong if it has more error cases. We only claim that is is less general in that its assumptions about \( u(c) \) disagree with the assumptions of the other six measures more often.

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 significant. We also use the G-test [25] to assure ourselves 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 incoherent 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 coefficient (labeled MCC for Matthews Correlation Coefficient) also does well, dispite measuring performance at only a single threshold.

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 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 incoherent 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.

A. Discussion

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 coherent point measure such as F1 or $\phi$ 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 coherence 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 coherent point measures, since 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 incoherent 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 $\phi$ would also seem to be a coherent and sufficiently general choice.

**B. A Note on Some Recent Related Work**

Our analyses, and those in [5], attempt to show that a certain class of performance measure does not measure the same quantity when presented with different classifiers tested on the same data, and is thus a broken tool for measuring performance. We do this by showing a linear relationship between this class of measures and a constructed measure that is also broken in this fashion. This relationship relies on the mapping between misclassification cost and threshold given in Equations 5 and 6.

Recent work [28] makes an important objection: While there is always a threshold that minimizes some cost, practitioners who know the cost will rarely find this optimal threshold in reality. As these analyses are based on this mapping, one may be able to dismiss them outright as having made assumptions that are unrealistic. Instead, the authors of [28] imagine an alternative way to view loss, referred to as $L^t_c$, based on separate integration over possible costs and thresholds:

$$L^t_c \triangleq \int_0^1 \int_0^\infty Q_c(t; b, c)W(t) \, dt \, w_c(c) \, dc$$

This gives us two choices to make, one for the distribution of thresholds, $W(t)$, and the other for the distribution of misclassification costs, $w(c)$. The authors show that the AUC is equivalent to choosing a uniform distribution for each. Thus, the AUC makes no assumptions about the cost distribution and coherence is restored. While the authors say nothing about other integrated measures, we do not doubt that the same argument may apply.

This is an important and valid argument against these analyses. Whom to believe? The choice here is between the analysis based on expected loss [28] or the one based on expected minimum loss [5]. At first, the argument based on expected loss may be more appealing on the grounds that it is clearly less optimistic. If our goal is to estimate absolute performance, this may be a virtue.

However, our primary goal here is comparison. The assumption in [28] is that all thresholds are equally likely even if the loss function is known. In reality, this is far from the truth: If a practitioner knows the loss function, she will be able to choose a reasonable threshold using cross-validation, a holdout set, or just intuition. She may not choose the optimal threshold, but she is likely to be closer (probably much closer) than random chance. In our analyses, we use the optimal threshold as a proxy for the practitioner’s choice. In [28], the proxy is a completely random guess. We leave it to the reader to decide which is a better basis for comparison.

It is also important to realize that even if one accepts the assumptions in [28], one may dismiss the AUC on the very grounds it is defended. The coherence of the AUC relies on the assumption all costs and all thresholds are equally likely. One may believe that the AUC is coherent on these grounds, but still reject the idea that the quantity it measures is inappropriate for classifier comparison. Our empirical analyses here would support such a rejection.

Finally, the authors show that the H-measure can be shown to be equivalent to using the “area under the cost curve” [29], with two modifications. One of these is that the H-measure assumes a beta distribution for $u(c)$ whereas cost curves assume a uniform distribution. The intuition for the use of
the beta distribution is that, first, it assumes a bias towards balanced costs given no other information, and second, the distribution can easily be modified to incorporate domain knowledge. The authors of [28] argue for the uniform distribution on the basis that “the cost lines are straight” and therefore easier to plot. We leave it to the reader to decide which is the stronger argument.

REFERENCES


