A Framework for Analyzing Skew in Evaluation Metrics

Alexander Liu **Joydeep Ghosh**

Department of Electrical & Computer Engineering
University of Texas at Austin, Austin, TX 78712, USA.
{aliu—ghosh}@ece.utexas.edu

Cheryl Martin

Applied Research Laboratories
University of Texas at Austin, Austin, TX 78713, USA.
cmartin@arlut.utexas.edu

Abstract

For several evaluation metrics for classification problems, correctly classifying an additional point from one class will have a different effect on the value of the evaluation metric compared to correctly classifying an additional point from another class. In this paper, we describe a method for quantifying these effects based on " metric skew". After describing how to find the skew for each class given a particular evaluation metric, we show what the skews are for several common evaluation metrics. In particular, we show that these skews provide a new viewpoint on metrics from which previously known as well as new properties about several popular metrics can be observed.

Introduction

Many evaluation metrics used to analyze the results in classification problems do not weight classifications from all classes equally. The most obvious example of this are evaluation metrics used in cost-sensitive learning which specifically take disparate misclassification costs into account. However, several metrics that do not use misclassification costs are also not symmetric with respect to the effect of correctly classifying points from different classes. Table **??** illustrates a simple example where this is true for a variety of common evaluation metrics. We look at the values of the metrics on two related two-class problems. In both cases, the number of points in the positive and negative classes are both equal to 100. In case 1, the number of true positives is 80 while the number of true negatives is 50; in case 2, the opposite is true: the number of true positives is 50 while the number of true negatives is 80. Note that only accuracy is equal in both cases, while for all other metrics, there is clearly a different effect due to misclassifications in the positive versus the negative class.

Several authors have noted that many evaluation metrics do not weight misclassification costs equally (e.g., (**?**)). In this paper, we introduce a framework for analyzing evaluation metrics that quantifies the impact of correctly classifying samples from each class. We then apply the framework to a number of common metrics. We show that our framework provides a new viewpoint on several known and previously unknown properties about metrics.

Table 1: A simple example of asymmetry

Metric	Case 1	Case 2
accuracy	0.65	0.65
precision	0.62	0.71
recall	0.80	0.50
f1-measure	0.70	0.59

A note on notation: We will use lower-case letters for scalars and bold lower-case letters for functions. $\frac{d\mathbf{f}}{dx}$ denotes the derivative of a function $\mathbf{f}$ with respect to x while $\frac{\partial \mathbf{g}}{\partial x}$ denotes the partial derivative of a function $\mathbf{g}$ with respect to x. For two-class problems, let constants n^+ and n^- represent the number of points in the positive and negative classes, respectively, and let $n = n^+ + n^-$ represent the total number of data points being evaluated. Let n_{tp} represent the current number of true positives and n_{tn} represent the current number of true negatives. Similarly, n_{fp} and n_{fn} will refer to the number of false positives and false negatives, respectively; note that, $n_{fn} = n^+ - n_{tp}$ and $n_{fp} = n^- - n_{tn}$, so any equations with n_{tp}, n_{tn}, n_{fp}, and n_{fn} can be written in terms of only n_{tp}, n_{tn}, and constants n, n^+, and n^- for a given problem. For the multi-class case, k represents the number of classes, n^j represents the number of points from the jth class, n^j_{tp} is the number of correctly classified points in class j and n^j_{fp} is the number of points incorrectly classified as class j.

Background on Evaluation Metrics

A number of popular evaluation metrics have been introduced in the past for analyzing the predictions of classification algorithms. Below, we describe several common metrics analyzed in this paper.

Total cost and variants

If misclassification costs are known, a useful evaluation metric is the total misclassification cost. Here, we discuss the case where there is a fixed misclassification cost for misclassifying a point from class i as a point from class j (**?**).

For a two-class problem, let the cost of making a false positive be denoted as c_{fp} and similarly for true positives

(c_{tp}), true negatives (c_{tn}), and false negatives (c_{fn}). Then the total cost is defined as [1]:

$$\mathbf{m_{totalcost}}(n_{tp}, n_{tn}) = c_{tp}n_{tp} + c_{tn}n_{tn} + c_{fn}n_{fn} + c_{fp}n_{fp} \quad (1)$$

In practice, c_{tp} and c_{tn} are often set to zero so as not to penalize the classifier for correct decisions. c_{fp} and c_{fn} are also normalized by c_{fp} such that costs c'_{fp} and c'_{fn} are used, where $c'_{fp} = 1$ and $c'_{fn} = \frac{c_{fn}}{c_{fp}}$. In this case, equation **??** can be rewritten as:

$$\mathbf{m'_{totalcost}}(n_{tp}, n_{tn}) = c'_{fn}n_{fn} + n_{fp} = c'_{fn}(n^+ - n_{tp}) + n^- - n_{tn} \quad (2)$$

Note that if equation **??** is divided by n, then we get the expected cost. In addition, if $c'_{fn} = 1$, then the expected cost becomes 1 minus the average accuracy of the classifier.

ROC and AUC

A ROC (receiver operating characteristic) curve can be thought of as a graphical metric used to evaluate classifiers. The horizontal axis is used to plot the false positive rate ($\frac{n_{fp}}{n_-}$) assigned by a particular classifier setting and the vertical axis is used to plot the true positive rate ($\frac{n_{tp}}{n_+}$) at the same classifier setting.

A convenient way of comparing ROC curves is the area under the curve (AUC), an aptly named numerical metric obtained by finding the area under the ROC curve. In practice, since only certain points on the ROC curve are found empirically, the AUC can be conveniently calculated using the trapezoid rule.

We define AUC as follows in order to analyze AUC for a single value of n_{tp} and n_{tn}. Note that the current values of n_{tp} and n_{tn} only affect one part of the ROC curve. Let (n_{tp1}, n_{tn1}) be the point on the ROC curve immediately to the left of (n_{tp}, n_{tn}) and let (n_{tp2}, n_{tn2}) be the point immediately to the right of (n_{tp}, n_{tn}) on the ROC curve. Then, using the trapezoid rule, the contribution to total AUC based on the current point (n_{tp}, n_{tn}) and its two immediate neighbors is given by:

$$\mathbf{m_{pAUC}}(n_{tp}, n_{tn}) = \frac{1}{2}\left(\frac{n_{tn1}}{n^-} - \frac{n_{tn}}{n^-}\right)\left(\frac{n_{tp}}{n^+} + \frac{n_{tp1}}{n^+}\right) + \frac{1}{2}\left(\frac{n_{tn}}{n^-} - \frac{n_{tn2}}{n^-}\right)\left(\frac{n_{tp}}{n^+} + \frac{n_{tp2}}{n^+}\right) \quad (3)$$

where the subscript "pAUC" is meant to highlight the fact that this only calculates part of the total AUC.

Note that the ROC curve always includes the two trivial classifiers that always predict the negative class or always predict the positive class. Thus, a trivial three-point ROC curve can be created for any single value of (n_{tn}, n_{tp}) by letting $(n_{tp1}, n_{tn1}) = (0, n^-)$ and $(n_{tp2}, n_{tn2}) = (n^+, 0)$. In addition, for the trivial three-point ROC curve, $\mathbf{m_{pAUC}}$ is equal to the AUC of the entire ROC curve.

[1] Note that defining cost in terms of n_{tp} and n_{tn} instead of n_{fn} and n_{fp} is somewhat unnatural; we do this such that all metrics are defined as functions of the same variables

Precision, recall, and variants

For a two-class problem, precision and recall are defined as follows:

$$\mathbf{m_{precision}}(n_{tp}, n_{tn}) = \frac{n_{tp}}{n_{tp} + n_{fp}} = \frac{n_{tp}}{n_- + n_{tp} - n_{tn}} \quad (4)$$

$$\mathbf{m_{recall}}(n_{tp}, n_{tn}) = \frac{n_{tp}}{n_{tp} + n_{fn}} = \frac{n_{tp}}{n_+} \quad (5)$$

Since it is convenient to look at a single metric instead of two separate metrics, precision and recall have been combined in a number of ways. One popular method is the f1-measure, the harmonic mean of precision and recall, which is equal to:

$$\mathbf{m_{f1-measure}}(n_{tp}, n_{tn}) = \frac{2 * n_{tp}}{n + n_{tp} - n_{tn}} \quad (6)$$

Another method of combining precision and recall is to take the geometric mean.

To extend precision and recall for multi-class problems, either the microaverage or macroaverage can be used. The microaveraged recall is defined as:

$$\mathbf{m_{mic.rec.}} = \frac{\sum_{j=1}^{k} n_{tp}^j}{\sum_{j=1}^{k} n^j} \quad (7)$$

while macroaveraged recall is defined as:

$$\mathbf{m_{mac.rec.}} = \frac{1}{k}\sum_{j=1}^{k} \frac{n_{tp}^j}{n^j} \quad (8)$$

Similarly, microaveraged precision is defined as:

$$\mathbf{m_{mic.prec.}} = \frac{\sum_{j=1}^{k} n_{tp}^j}{\sum_{j=1}^{k} n_{tp}^j + \sum_{j=1}^{k} n_{fp}^j} \quad (9)$$

while the macroaveraged precision is defined as:

$$\mathbf{m_{mac.prec.}} = \frac{1}{k}\sum_{j=1}^{k} \frac{n_{tp}^j}{n_{tp}^j + n_{fp}^j} \quad (10)$$

Metric Skew

In this section, we present our approach for determining the effect of classifying an additional point from a specific class on the value of an evaluation metric. The method is straightforward and simple, and involves finding the derivative of a metric $\mathbf{m}$ with respect to changes in the number of correctly classified instances from that class.

Let us begin by looking at the simplest possible case: a two-class problem. In order to determine how much the metric will change if we add an additional h true positives or h true negatives, we can define the following delta functions:

$$\boldsymbol{\delta}_+(n_{tp}, n_{tn}, h) = \frac{\mathbf{m}(n_{tp} + h, n_{tn}) - \mathbf{m}(n_{tp}, n_{tn})}{h} \quad (11)$$

and

$$\boldsymbol{\delta}_{-}(n_{tp}, n_{tn}, h) = \frac{\mathbf{m}(n_{tp}, n_{tn} + h) - \mathbf{m}(n_{tp}, n_{tn})}{h} \quad (12)$$

where $\mathbf{m}(n_{tp}, n_{tn})$ is the value of some evaluation metric $\mathbf{m}$ given n_{tp} and n_{tn}. Thus, a metric will change by $\boldsymbol{\delta}_{+}(n_{tp}, n_{tn}, 1)$ if the number of true positives is increased by 1 and will change by $\boldsymbol{\delta}_{-}(n_{tp}, n_{tn}, 1)$ if the number of true negatives is increased by 1.

If we take the limit of $\boldsymbol{\delta}_{+}(n_{tp}, n_{tn}, h)$ as h approaches 0, we get the derivative of the metric with respect to n_{tp}. We denote this derivative as $\frac{\partial \mathbf{m}}{\partial n_{tp}}$. Similarly, $\frac{\partial \mathbf{m}}{\partial n_{tn}}$ is equal to the limit of $\boldsymbol{\delta}_{-}(n_{tp}, n_{tn}, h)$ as h approaches 0. Note that this method can be easily extended to the general multi-class case by computing the derivatives of the metric with respect to the number of points correctly classified in each class.

Let us define the skew of a metric towards the jth class as $\frac{\partial \mathbf{m}}{\partial n_{tp}^{j}}$. For the special two-class case, we call the derivatives $\frac{\partial \mathbf{m}}{\partial n_{tp}}$ and $\frac{\partial \mathbf{m}}{\partial n_{tn}}$ the metric skew towards the positive and negative classes, respectively. We use the term "skew" for a number of reasons. The first is to avoid confusion with other potential names, such as cost or bias, which are already commonly used for other purposes. The second is that the term "skew" has the connotation of a slant or bias towards a particular direction. A third reason is its connection with a quantity known as the effective skew ratio (**?**) which we will describe in more detail below.

First, however, let us define the skew ratio of a metric for two-class problems. **Let the skew ratio be defined as the skew towards the positive class divided by the skew towards the negative class (i.e., skew ratio$= \frac{\partial \mathbf{m}/\partial \mathbf{n_{tp}}}{\partial \mathbf{m}/\partial \mathbf{n_{tn}}}$).** If the skew ratio is greater than one, then increasing the number of true positives will increase the value of the metric more than increasing the number of true negatives; that is, the metric is skewed towards the positive class. The skew ratio is a convenient quantity when looking at metric skews in two-class problems. However, while metric skews can be defined for multi-class problems, the skew ratio cannot. For multi-class problems, we must compare individual metric skews against each other (we will show an example of this for micro/macro averaged precision/recall).

In (**?**), a quantity known as the effective skew ratio was defined. The effective skew ratio is the slope of a metric's isometric lines in ROC space and is related to the skew ratio by the following theorem:

Theorem 0.1. *For a two-class problem, the effective skew ratio of a metric is equal to $\frac{n^-}{n^+}$ times the inverse of the skew ratio.*

Proof. Denote a metric in ROC space as a function $\mathbf{m_{ROC}}(n_{tpr}, n_{fpr})$ where the true positive rate is defined as $n_{tpr} = \frac{n_{tp}}{n^+}$ and the false positive rate is defined as $n_{fpr} = \frac{n_{fp}}{n^-} = \frac{n^- - n_{tn}}{n^-}$. $\mathbf{m_{ROC}}(n_{tpr}, n_{fpr}) = \mathbf{m}(n^+ n_{tpr}, n^- - n^- n_{fpr})$ where $\mathbf{m}$ is a metric defined as some function of n_{tp} and n_{tn} as done previously in this paper.

An isometric line of a metric in ROC space is defined by letting $\mathbf{m_{ROC}}(n_{tpr}, n_{fpr}) = m'$ where m' is some constant. The effective skew ratio is the slope of this line in ROC space and is equal to $\frac{dn_{tpr}}{dn_{fpr}}$.

In calculus, implicit differentiation states the following: If a differentiable function $\mathbf{f}(x, y) = 0$, where y is defined implicitly as a differentiable function of x, and $\frac{\partial \mathbf{f}}{\partial y} \neq 0$, then $\frac{dy}{dx} = -\frac{\partial \mathbf{f}/\partial x}{\partial \mathbf{f}/\partial y}$.

Let $\mathbf{f}(n_{tpr}, n_{fpr}) = \mathbf{m_{ROC}}(n_{tpr}, n_{fpr}) - m' = 0$. Then, because of implicit differentiation, the effective skew ratio $\frac{dn_{tpr}}{dn_{fpr}}$ is equal to $-\frac{\partial \mathbf{f}/\partial n_{fpr}}{\partial \mathbf{f}/\partial n_{tpr}}$.

Note that: $\frac{\partial \mathbf{f}}{\partial n_{tpr}} = \frac{\partial \mathbf{m_{ROC}}}{\partial n_{tpr}} = \frac{\partial \mathbf{m}}{\partial n_{tp}} \frac{dn_{tp}}{dn_{tpr}} = n^+ \frac{\partial \mathbf{m}}{\partial n_{tp}}$ and $\frac{\partial \mathbf{f}}{\partial n_{fpr}} = \frac{\partial \mathbf{m_{ROC}}}{\partial n_{fpr}} = \frac{\partial \mathbf{m}}{\partial n_{tn}} \frac{dn_{tn}}{dn_{fpr}} = -n^- \frac{\partial \mathbf{m}}{\partial n_{tn}}$

Thus, effective skew ratio $= \frac{dtpr}{dfpr} = -\frac{\partial \mathbf{f}/\partial fpr}{\partial \mathbf{f}/\partial tpr} = \frac{n^-}{n^+} \frac{\partial \mathbf{m}/\partial n_{tn}}{\partial \mathbf{m}/\partial n_{tp}} = \frac{n^-}{n^+} \frac{1}{\text{skew ratio}}$. □

Both the effective skew ratio and metric skew can be used for analyzing and characterizing metrics. However, using metric skew has several advantages. First, since determining the effective skew ratio depends on isometrics in ROC space, it is more difficult to extend beyond two-class problems (we are unaware of any such extensions). However, metric skews can be calculated for each class in a multi-class problem as long as the metric is defined in terms of constants and n_{tp}^{j} for all classes j. Furthermore, the method of calculating metric skew is more straightforward than the example method of finding effective skew ratios based on isometrics in ROC space as described in (**?**). Finally, for two-class problems specifically, using the ratio of the two metric skews for each class (the skew ratio) rather than effective skew ratio is better suited for determining whether a metric is biased towards the positive or negative class. For example, if the skew ratio is exactly equal to 1, then the metric is not skewed towards either class. However, a skew ratio of 1 means that the metric has an effective skew ratio of $\frac{n^-}{n^+}$. Here, the effects of classifying the positive and negative class is less straightforward to interpret.

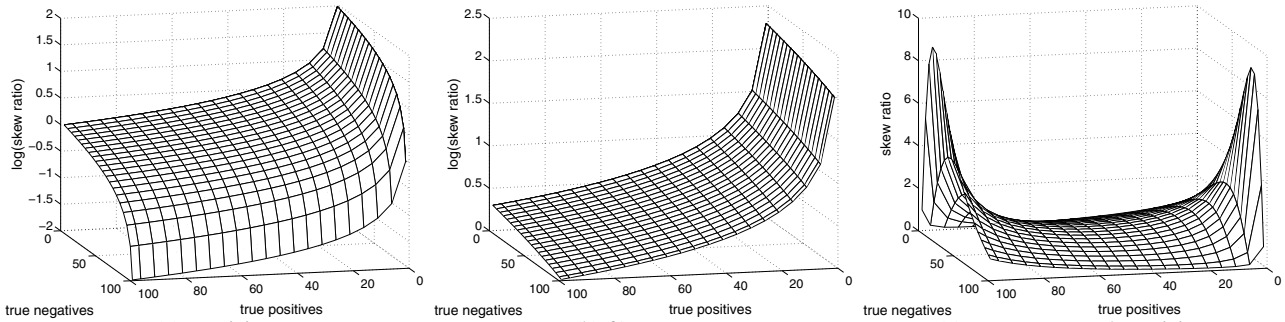
Analysis of Common Metrics

Let us now use metric skew to determine the skew ratios of several common metrics in two-class problems. We will also use metric skew to examine some example metrics for multi-class problems. We will see that our approach not only confirms and is consistent with known observations on evaluation metrics, but also provides several new properties.

A summary of skew ratios for two-class problems are in table **??**. Figure **??** plots some of the non-constant skew ratios from table **??**. In all plots in figure **??**, $n^+ = n^- = 100$ while n_{tp} and n_{tn} are varied. Note that the plots show the log of the skew ratio for precision and f1-measure, but not macroaveraged precision.

Total cost and variants

First, let us show that our method of finding metric skew is consistent with cost-sensitive metrics. Using equation **??** as the evaluation metric, $\frac{\partial \mathbf{m}}{\partial n_{tp}} = -c'_{fn}$ and $\frac{\partial \mathbf{m}}{\partial n_{tn}} = -1$. Thus, the skew ratio equals c'_{fn}, meaning that misclassifying the



(a) precision (b) f1-measure (c) macroaveraged precision

Figure 1: Example graphs of non-constant skew ratios when $n^+ = 100$, $n^- = 100$

positive class is c'_{fn} times more costly than misclassifying the negative class, a result that is exactly consistent with the definition of costs in cost-sensitive learning. Similarly, it is easy to show that the skew ratio of average accuracy is equal to 1, which is consistent with the fact that true positives and true negatives contribute equally to average accuracy.

AUC

For AUC, we get a particularly interesting skew ratio. Using equation **??** (which only calculates the contribution of the current values of n_{tp} and n_{tn} to the total AUC based on its immediate neighbors), we obtain a skew ratio of $\frac{n_{tn1}-n_{tn2}}{n_{tp2}-n_{tp1}}$, meaning that the relative importance of correctly classifying the positive versus the negative class depends strictly on the values in the ROC curve to the immediate left and right of the current point. As discussed, a single value of n_{tp} and n_{tn} returned by a classifier can be turned into a three-point ROC curve by including the trivial classifier that always guesses the negative class and the classifier that always guesses the positive class. In this special case, the skew ratio is equal to $\frac{n^-}{n^+}$, meaning that correct classifications from each class are weighted proportionally to the inverse of its prior.

Precision, recall, and variants

For recall, the skew for the positive class is $1/n^+$ and the skew for the negative class is 0. Since recall depends only on performance on the positive class, it is always more important to do well on the positive class, which is consistent with the values for the metric skews. The skew ratio of precision is $\frac{n^- - n_{tn}}{n_{tp}}$; the log of this skew ratio is plotted in Fig. **??**(a).

For f1-measure, the skew ratio is equal to $\frac{n-n_{tn}}{n_{tp}}$. Since $n \geq n_{tn} + n_{tp}$, the skew ratio is always greater than or equal to 1, with equality only if there are no false positives and no false negatives. At this point, of course, there is no way to increase f1-measure, meaning that increasing the number of true positives always has a greater effect than increasing the number of true negatives. In Fig. **??**(b), the log of the skew ratio of f1-measure is plotted. Note that the skew ratio is largest when there are very few true positives and smallest when n_{tp} and n_{tn} are close to n^+ and n^-.

The skew ratio of the geometric mean of precision and recall is equal to $\frac{2n^- + n_{tp} - 2n_{tn}}{n_{tp}}$. Since the maximum value of n_{tn} is n^-, the skew ratio is also always greater than or equal to 1.

Analysis of metrics for multi-class problems

Now let us use the microaveraged and macroaveraged precision and recall as examples of how our approach can be applied to the more general multi-class case. In order for a more direct comparison with our previous analysis on two-class metrics, the skew ratios of these metrics are also described below and included in table **??**.

The skew of microaveraged recall for the jth class is $\frac{1}{n}$. Since this is a constant, the skew of microaveraged recall for all classes is equal. Thus, all classifications contribute equally to microaveraged recall, and, for the two-class case, the skew ratio is 1. The skew of macroaveraged recall for the jth class is $\frac{1}{n^j}$. That is, classifications are weighted with weights equal to the inverse of the number of points in that class. In the two-class case, this means the skew ratio is $\frac{n^-}{n^+}$. This is consistent with previous work.

The skew of the microaveraged precision is also a constant, meaning that all classifications contribute equally to microaveraged precision. However, the skew of macroaveraged precision is not the same as the skew of macroaveraged recall and is much less straightforward. For the simplest two-class case, the skew ratio of macroaverage precision is equal to:

$$\frac{n^-(n^+ + n_{tn} - n_{tp})^2 + n_{tn}(n)[n^- - n^+ + 2(n_{tp} - n_{tn})]}{n^+(n^- + n_{tp} - n_{tn})^2 + n_{tp}(n)[n^+ - n^- + 2(n_{tn} - n_{tp})]}$$

which is in general not equal to $\frac{n^-}{n^+}$. This skew ratio is plotted in Fig. **??**(c) for the case where $n^+ = n^- = 100$. As evident from the figure, the skew ratio of macroaverage precision is a saddle function. Thus, macroaveraged precision does not weight classifications inversely proportional to the number of points from that class in the test set. Instead,

Table 2: A summary of evaluation metrics and their skew ratios

Metric	Skew ratio
accuracy	1
total cost	c'_{fn}
partial AUC	$\frac{n_{tn1}-n_{tn2}}{n_{tp2}-n_{tp1}}$
recall	see discussion
precision	$\frac{n^- - n_{tn}}{n_{tp}}$
f1-measure	$\frac{n - n_{tn}}{n_{tp}}$
geo. mean of prec. and rec.	$\frac{2n^- + n_{tp} - 2n_{tn}}{n_{tp}}$
microaverage recall	1
macroaverage recall	$\frac{n^-}{n^+}$
microaverage precision	1
macroaverage precision	see discussion

macroaveraged precision weights classifications very differently depending on the current number of correct classifications.

Discussion

Our work shows that care must be taken when misclassification costs are either equal or unknown. If misclassification costs are equal, then care must be exercised when using and interpreting results from a metric with a skew ratio that is not 1. For example, in a two-class problem, precision, recall, and f-measure are often used in cases where both classes are equally important; however, this might lead to incorrect conclusions. In cases where classes are equally important, the *microaveraged* precision, recall, and f-measure should be used instead.

In cases where misclassification costs are unequal and unknown, the use of certain evaluation metrics with certain values of skew results in an implicit misclassification cost for each class that is equal to the skew for that class. An example where misclassification costs are unequal and unknown is the imbalanced dataset problem.

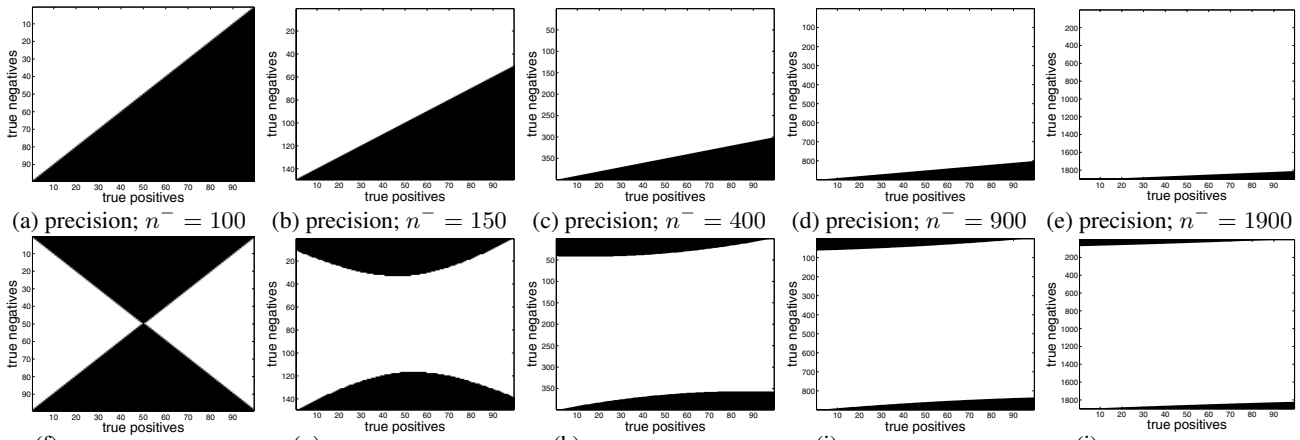
In the imbalanced dataset problem, the priors of each class are highly unequal. For example, in a two-class case, the probability of a datapoint from the positive class is much smaller than the probability of a datapoint from the negative class. Without accounting for class imbalance, many classifiers often learn to assign all points to the negative class. The argument is that this type of classifier is well-nigh useless, and, in example domains such as cancer detection, various types of fraud detection, and intrusion detection, this argument holds true.

However, if misclassification costs are equal, a classifier that learns to always predict the class with the highest prior may, in fact, be a very good learned classifier. For example, if misclassification costs are equal and the prior probability of the negative class is, say, 99%, then a classifier that always assigns points to the negative class is just as valid as a classifier which is completely correct on the positive class but misclassifies 1% of the data points (i.e., 1/99th of the negative class). While the second classifier seems more palatable, the total number of misclassifications is the same in both cases [2]. In practice, it may also be difficult to increase the number of true positives without introducing a large number of false positives as well, meaning that a classifier which always guesses the negative class may in fact lead to the lowest number of total misclassifications. Thus, a classifier learned from an imbalanced dataset which always predicts a single class is a problem only when misclassifying the minority class has some cost that is higher and unequal to the cost of misclassifying the majority class.

Thus, the imbalanced dataset problem involves both unequal priors as well as a (possibly unknown) higher misclassification cost for the minority class. Many past papers on imbalanced datasets have stated that an evaluation metric such as total accuracy or total number of misclassifications is susceptible to imbalanced class priors; instead, metrics such as AUC and f1-measure have been used because of their relative "immunity" to imbalanced class priors. As we have argued, application of these metrics to a problem where costs are supposedly "unknown" implies that hidden but known costs (equivalent to metric skew) are being used during evaluation. Neither AUC nor f1-measure have a skew ratio guaranteed to equal 1. F1-measure always has a higher skew for the positive class. For AUC, when $n^+ < n^-$ as in imbalanced datasets, the skew ratio may often be greater than 1. For example, in the simple three-point ROC curve, the skew ratio is always equal to $\frac{n^-}{n^+}$ which, in an imbalanced dataset, is always greater than 1. Thus, the ability of AUC and f1-measure to effectively rank classifiers on imbalanced datasets seems to stem partially from the fact that the skew ratio favors the minority class, a bias which matches the property that the minority class has an implicit but unspecified misclassification cost greater than the misclassification cost on the majority class.

Other metrics also become more skewed towards the positive class as the prior of the negative class becomes larger than the prior of the positive class. In figure **??**, we graph the indicator functions of whether the skew ratio is greater than 1; in the graphs, the color white indicates that the skew ratio is greater than 1 for those values of n_{tp} and n_{tn}, while black indicates that the skew ratio is less than or equal to 1. In the top row, we plot precision, while in the bottom row, we plot macroaveraged precision. The size of the negative class becomes larger with respect to the size of the positive class as one goes from left to right in the figure. Note that as the relative size of the negative class increases, the relative amount of space in the figures where the skew ratio is biased towards the positive class becomes larger and larger. In addition, only the two right-most columns have a positive class prior less than or equal to 0.1. That is, even when the negative class is slightly larger with respect to the positive class, the skew ratio favors the positive class for most possible values of n_{tp} and n_{tn}.

[2] This is related to the problem with accuracy described in (**?**); using accuracy as an objective measure, however, each of these classifiers is equally good



(a) precision; $n^- = 100$ (b) precision; $n^- = 150$ (c) precision; $n^- = 400$ (d) precision; $n^- = 900$ (e) precision; $n^- = 1900$

(f) macroave. prec.; $n^- = 100$ (g) macroave. prec.; $n^- = 150$ (h) macroave. prec.; $n^- = 400$ (i) macroave. prec.; $n^- = 900$ (j) macroave. prec.; $n^- = 1900$

Figure 2: Graphs of indicator function for whether skew ratio> 1 for $n^+ = 100$ and for different values n^-. On the graph, white indicates skew ratio> 1 and black indicates skew ratio$<= 1$. Note that even for moderate amounts of imbalance, both example metrics are skewed towards the positive class for most possible values of n_{tp} and n_{tn}.

Future Work

One area of future work is to further investigate non-constant metric skews. In particular, it is less obvious how to use non-constant metric skews in a beneficial manner in comparison to constant metric skews. Another area of future work is to use the metric skew and skew ratios to directly compare metrics. Any method of using metric skew to compare evaluation metrics will need to take prior work on this area in (**?**) and (**?**) into consideration. We also plan to further investigate the relationship between metric skew and misclassification costs in cost-sensitive learning. Finally, one possibility is to use the metric skews themselves to weight classification results in order to handle problems such as dataset imbalance and to compare such an approach with ROC curves and cost curves (**?**).

Summary and Conclusion

In this paper, we have introduced a method for finding the skew of evaluation metrics. To the best of our knowledge, the only other comparable approach is the one described in (**?**) for finding effective skew ratios. However, our method is simpler to apply, easier to interpret, and directly applicable to metrics used on multi-class problems. It is also straightforward, consistent with known facts about metrics, and useful for finding previously unknown facts about metrics. Among the known facts that can be directly derived from our approach are the following:

1. accuracy weighs all classifications equally
2. recall depends only on classifications on the positive class
3. microaveraged recall and precision weigh all classifications equally
4. macroaveraged recall weighs classifications with a weight inversely proportional to the number of points in that class.

More importantly, previously unknown facts about metrics presented in this paper include:

1. the exact form for the metric skew of macroaveraged precision
2. f1-measure and the geometric mean of precision and recall are always skewed towards the positive class
3. AUC may be skewed towards the positive class in imbalanced problems.

Acknowledgments: Thanks to anonymous reviewers, members of IDEAL, and members of CIADS laboratories for helpful comments. We would particularly like to thank Dung Lam whose insightful comments motivated this work.

References

Drummond, C., and Holte, R. C. 2004. What ROC curves can't do (and cost curves can). In *ROCAI-2004*, 19–26.

Elkan, C. 2001. The foundations of cost-sensitive learning. *Proc. 17th IJCAI* 973–978.

Flach, P. A. 2003. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *Proc. 20th ICML*, 194–201.

Japkowicz, N. 2006. Why question machine learning evaluation methods? *AAAI-06 Evaluation Methods for Machine Learning Workshop* 6–11.

Sokolova, M.; Japkowicz, N.; and Szpakowicz, S. 2006. Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. In *Australian Conference on Artificial Intelligence*, 1015–1021.

Vilalta, R., and Oblinger, D. 2000. A quantification of distance-bias between evaluation metrics in classification. In *Proc. 17th ICML*, 1087–1094.