On Evaluating Performance of Classifiers for Rare Classes

Mahesh V. Joshi

IBM T. J. Watson Research Center

P.O.Box 704, Yorktown Heights, NY 10598

joshim@us.ibm.com

Abstract

Predicting rare classes effectively is an important problem in many critical domains. The definition of effective *classifier is however very subjective, primarily dependent on the application domain of the classifier. In this paper, we address the issue of choosing the right metric in a more domain-independent and objective manner. We compare a wide variety of metrics: simple ones derived using the confusion matrix as well as the complex information theoretic metrics. There are two key contributions. First, we put all the metrics into a common analytical context defined by the recall and precision of the target rare class. This allows us to make objective comparisons. Secondly, we judge their suitability for the rare class problems over wide variations of learning difficulty and levels of rarity. This yields many valuable insights regarding the equivalency of some metrics, each one's preference for recall over precision, and choice of a more generally suitable metric. In order to reduce the dependency on one single composite metric, we propose a way of comparing classifiers that weighs the actual relationship between recall and precision, and looks at a composite metric only when individual metrics yield conflicting results. We validate our analysis and arguments using various real-world and synthetic datasets.*

1 Introduction

Learning effective predictive models for rarely occurring events is an important problem with applications in many domains such as document categorization, network intrusion detection, web-log analysis, etc. Recently, new algorithms and insights have started appearing with respect to this problem [6, 4, 8, 7]. However, the issue of evaluating the classifiers for these problems has not been properly addressed in a comprehensive, domain-independent manner. Various metrics exist in the literature to evaluate classifiers in general, primarily motivated by the domain in which they are applied [14, 2, 12, 15, 11, 5, 9, 16].

Our primary goal in this paper is to give a common objective perspective to various such metrics ranging from the traditional accuracy metric to a complex information theoretic metric [9]. An attempt is made to judge their suitability for evaluating the performance of the rare class prediction problem. We take a stand that for effective prediction a rare class, two metrics are crucial: recall and precision *with respect to* to the rare class[1]; i.e. it is important to predict as many of the future occurrences of rare class as possible at the same time maintaining a low ratio of incorrect predictions (false positives) to the correct predictions. We translate many different metrics into this recall-precision perspective and evaluate their comparative relationships for varying levels of rarity of the target class and various difficulties of learning problems defined with respect to the recall and precision levels.

Although no single metric is best suited for all scenarios, the insight that we gain through our analysis is that some metrics are better suited than others especially for the rare class problems because of their consistent performance over various ranges of rarity values and recall and precision levels. Also, this fact leads us to propose a way of comparing classifiers that is not entirely dependent on one single composite metric. It compares the core individual metrics of recall and precision, and uses the composite metric only to resolve conflicts.

The rest of the paper is organized as follows. First we define various metrics in Section 2. Then we compare them analytically from a common perspective in Section 3 and empirically in Section 4. Various real-world and synthetic data-sets are used to validate the analytically derived relationships between various metrics. Section 6 summarizes the insights gained from the analysis.

2 Evaluation Metrics

We focus on the binary classification problem involving two classes: C and NC. Different metrics have been proposed in the literature to assess the classifier performance when one of the classes is rare. In the context of binary problems, most of the metrics can be derived using a *confusion matrix*. Let us assume that the classifier is being evaluated over a dataset T of n records. The confusion matrix of Table 1 records the distribution of examples of T into four types based on their predicted and actual classes.

The total number of class C examples in T is $n^c = n_{00} + n_{01}$ and total number of NC examples in T is $n^{nc} = n_{10} + n_{11}$. TP are true positives, FP the false positives, TN the true negatives, and FN the false negatives.

[1]Given a binary classification problem of learning a model to distinguish a rare class C from the other class (NC), if a classifier is able to predict m cases of C correctly out of total n cases, then the Recall (R) of the classifier with respect to C is m/n. If the classifier makes total l predictions as class C (out of which only m are correct), then the precision with respect to C is m/l.

	Predicted as C	Predicted as NC	Total
Actually C	n_{00} [TP]	n_{01} [FN]	n^c
Actually NC	n_{10} [FP]	n_{11} [TN]	n^{nc}
Total	$n_0 = n_{00} + n_{10}$	$n_1 = n_{01} + n_{11}$	n

Table 1. Confusion Matrix for a Binary Classification Problem

The terms "positive" and "negative" are defined with respect to C. The false negatives are often called Type I errors, and false positives are often called Type II errors.

Many different combinations of the values n_{ij} can be used to form different classifier evaluation metrics. Here are the definitions of some core metrics that we will use. Recall and Precision of the classifier with respect to a class c, are denoted by R^c and P^c, respectively. R^c measures how many examples of class c are correctly predicted by the classifier, and P^c measures how many examples predicted as class c by the classifier are indeed of class c. Using the notations of our confusion matrix,

$$R^c = n_{00}/(n_{00} + n_{01}) \quad (1)$$

$$P^c = n_{00}/(n_{00} + n_{10}) \quad (2)$$

$$R^{nc} = n_{11}/(n_{11} + n_{10}) \quad (3)$$

$$P^{nc} = n_{11}/(n_{11} + n_{01}) \quad (4)$$

Before listing various evaluation metrics that can be derived from the confusion matrix and have been used in the past, we would like to put the matter of judging the classifier performance in perspective. In practice, there usually is a tradeoff between achieving lower type I errors versus achieving lower type II errors. It is very difficult to simultaneously achieve perfect values (0) for both, except when the two classes are completely uncorrelated in a noise-free environment. Also, a classifier can usually be tuned to yield more or less of one type of the error by varying some of its parameters. For a classifier which assigns a probabilistic score to each of its prediction, this tuning can be done by varying a score threshold at which the class prediction changes (i.e. varying the value of v such that a score $< v$ implies class NC and score $\geq v$ implies class C).

Realizing this, one can plot what is called a relative operating curve (ROC) characteristics of a classifier [14], which is inspired by signal detection theory [3]. The ROC curve records various combinations type I and II errors obtained by a classifier algorithm on a given problem by varying its parameters. One definition of ROC curve is the plot of R^c versus $(1 - R^{nc})$ values. An example of a typical such curve is given in Figure 1. R^{nc} is just one

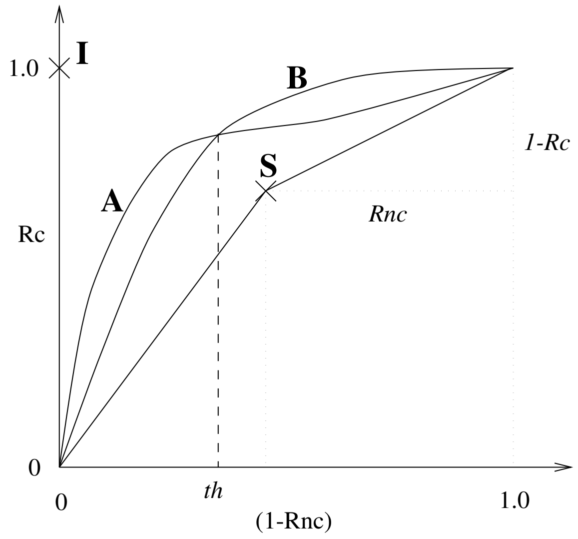

Figure 1. ROC curves of R^c vs. $(1 - R^{nc})$. Classifier **A** is better (resp. worse) than classifier **C** (resp. better) for $(1 - R^{nc}) < th$ (resp. $(1 - R^{nc}) > th$). Classifier **I** (top left corner) is the best possible. Also illustrates AUC of single classifier, **S**. Similar ROC curves can be drawn for R^c vs. $(1 - P^c)$.

way of accounting for the type II errors; i.e., as a fraction of the actual NC examples. Precision of C accounts for type II errors as a fraction of total C predictions. Hence, ROC curve can also be defined using recall and precision achieved for C; i.e. by plotting R^c vs. $(1 - P^c)$.

Two classifiers can be compared by formulating some *whole-curve metric* that quantifies the ROC curve as a whole. One such metric is the area-under-the curve (AUC) [2]. A classifier that can achieve 100% recall for both the classes simultaneously is the best possible classifier (denoted by **I** in Figure 1), and has an AUC of 1. On two extremes, 100% R^c can be achieved with 0% R^{nc} by predicting everything to be of C, or 100% R^{nc} and 0% R^c by predicting everything to be of NC. An interesting point to note is that AUC of a single classifier (for which ROC curve consists of the two extreme points and the point for that classifier's performance), is simply the arithmetic mean of R^c and R^{nc}, when ROC curve is defined as the R^c vs. $1 - R^{nc}$ curve. This is illustrated in Figure 1 for classifier **S**, for which $AUC = R^c \times (1 - R^{nc})/2 + R^c \times R^{nc} + R^{nc} \times (1 - R^c)/2 = (R^c + R^{nc})/2$. Similarly, if ROC curve is the R^c vs. $(1 - P^c)$ curve, then AUC of a single classifier will be $(R^c + P^c)/2$.

Whichever way the ROC curve is plotted, comparisons based on the whole curve are suitable only if one is looking for a classifier that performs the best across the entire range of achievable recall values of a given class (*best* implies that the ROC curve of a classifier lies entirely above the ROC curve of any other classifier). More often than not, no single classifier can perform the best across all the range. A typical behavior is illustrated in

Figure 1 by classifiers **A** and **B**. **A** is better than **B** when $R^{nc} > 1 - th$ and is worse than **B** when $R^{nc} < 1 - th$. An attempt has been made to form a hybrid classifier that can perform best throughout by combining multiple classifier techniques each of which is optimal only over certain range of the curve [13].

A classifier that performs best throughout the range of ROC (such as the hybrid classifier of [13]) is especially desired when the population on which the classifier is to be tested is expected to have significantly different class distributions and cost factors (on type I and I errors) than those assumed while training the classifier. However, the underlying assumption of this paper, just like most machine learning algorithms [12], is that the training and test samples are drawn from the same population in a manner that preserves class distributions. In fact, with no knowledge of the test population, this is the best assumption to make; otherwise, no single learning algorithm can be tested reliably and effectively. Under this assumption, we are interested in comparing classifier algorithms according to their ability to achieve the best single point on the ROC curve according to some *point-metric* versus the metric that takes the whole curve into consideration. Essentially, a point metric is a metric derived from the confusion matrix.

We first list different point-metrics that are commonly used in different communities.

- **Accuracy (Acc)**: This metric is widely used in machine learning [12]. It is defined as $(n_{00} + n_{11})/n$. For rare class problems, accuracy is not a desired metric, because for a problem involving 1% population of class C, predicting everything to be of NC will achieve 99% accuracy, which is highly acceptable from global perspective. But, from the perspective of effectively predicting the rare class, it is completely unacceptable.

- **F-measure (F)**: This is a metric used widely by the Information Retrieval community [15]. It is defined with respect to a given class. The generic form of F with respect to a class c is:

$$F^c_\lambda = \frac{1}{\lambda \, \frac{1}{R^c} + (1-\lambda) \, \frac{1}{P^c}}, 0 \leq \lambda \leq 1.$$

  We are interested in the classifier performance for the rare class; hence, we use F^c_λ as the metric. The value of λ can be selected depending on how much weight one wants to assign to reducing false negative predictions (i.e. increasing recall) versus reducing false positive prediction (i.e. increasing precision). This is usually problem-dependent, and can be decided if one knows the costs of each types of errors. For $\lambda = 0.5$, recall and precision get equal weight. Increasing λ gives more weight to recall and decreasing it emphasizes precision. With no a-priori knowledge of costs associated with either type of error, we use $F^c_{0.5}$ to evaluate the classifier. Henceforth, we refer to it simply as $F^c = 2\, R^c\, P^c/(R^c + P^c)$. Noted that this expression for F^c is directly proportional to the harmonic mean of R^c an P^c. This definition of F for

$\lambda = 0.5$ is also referred to as F_1 by the IR community, because usually λ is formulated as $\lambda = 1/(1+\beta^2)$, and F is parameterized on β ($\beta = 1$ gives $\lambda = 0.5$).

- **Geometric Mean of Recall and Precision (GMRP):** This metric is used by some in the information retrieval community [11]. It is also defined with respect to a particular class. For C, GMRP is $\sqrt{R^c\,P^c}$.
- **Sum of Recalls (SOR)**: This metric is used by some [5], and is also sometimes referred to as the weighted accuracy. It is simply what it says: SOR $= R^c + R^{nc}$.
- **Geometric Mean of Recalls (GMOR)**: This is used by some researchers for the rare class problem [11]. GMOR is defined as $\sqrt{R^c\,R^{nc}}$.
- **Information Score (IS)**: A metric inspired from information-theoretic perspective is proposed in [9]. The measure intends to cater to the rare class problems. It is a unique metric, in the sense that it attempts to reward the correct classification of a minority class example more than it rewards correct classification of a majority class example; i.e., in terms of our confusion matrix, it differentiates between true positives and true negatives by rewarding the former more. Moreover, it attempts to penalize the misclassification of a prevalent class example more than the misclassification of a rare class example; i.e., it emphasizes on achieving more precision. Given class of an instance C, and its prior probability $P(C)$, if the classifier predicts the class C for that example with probability $P'(C)$, then the information score, I for that instance is defined as:

$$
\begin{aligned}
I &= -\log P(C) + \log P'(C), \text{if } P'(C) \geq P(C) \\
&= \log(1 - P(C)) + \log(1 - P'(C)), P'(C) < P(C)
\end{aligned}
$$

I essentially tries to capture the amount of information (captured by the entropy) obtained by the prediction of C with probability $P'(C)$. An aggregate information score for a test set is obtained by computing I_a, the average information score, defined as $I_a = (\sum_{j=1}^{n} I(j))/n$.

A simplistic way to compute I_a for our confusion matrix is to assume that the classifier makes true and false predictions of C with 100% probability (if the classifier returns a probabilistic score for each prediction, then we can apply certain pre-determined threshold to the scores, and then regard everything above the threshold to have 100% score and everything below to have 0% score). Thus, for all the instances predicted as C; i.e., for true and false positives, $P'(C) = 1$ and $P'(NC) = 0$. Similarly, for true and false negatives, $P'(C) = 0$ and $P'(NC) = 1$. The prior probability of class C is the same as depicted by the confusion matrix, which fits well with our assumption that the training and test samples are drawn from same population. Hence, $P(C) = R^c$ and $P(NC) = R^{nc}$. Using these values, I value for each true positive prediction is $I^{TP} = -\log(R^c) + \log(1)$; for each false positive prediction, $I^{FP} = \log(1 - R^{nc}) + log(1 - 0)$; for each true negative prediction, $I^{TN} = -\log(R^{nc}) + log(1)$; and for each false negative prediction, $I^{FP} = \log(1 - R^c) + log(1 - 0)$. Thus, I_a can be expressed as,

$$\begin{aligned} I_a &= \frac{n_{00}}{n} I^{TP} + \frac{n_{01}}{n} I^{FN} + \frac{n_{10}}{n} I^{FP} - \frac{n_{11}}{n} I^{TN} \\ &= -\frac{n_{00}}{n} \log(R^c) + \frac{n_{01}}{n} \log(1 - R^c) + \\ &\quad \frac{n_{10}}{n} \log(1 - R^{nc}) - \frac{n_{11}}{n} \log(R^{nc}) \end{aligned} \tag{5}$$

- **Break-even Point**: The information retrieval community sometimes uses this metric to measure the level at which recall and precision can be made equal. It is based on the realization that recall and precision have a tradeoff. For a classifier that assigns probabilistic scores to its decision, by adjusting the threshold on the scores, one can obtain higher or lower values of one and accordingly lower or higher values for the other. Higher the level at which recall and precision can be equalized, the better the classifier. However, often this level has to be computed by interpolation, because it is difficult to realize a threshold at which recall and precision are equal on a real test dataset. Moreover, in cases where there is no probabilistic output from the classifier, this metric cannot be computed. Also, the flavor of this metric is captured by F^c metric defined earlier, because F^c is maximized when both R^c and P^c are equal.

3 Analytical Comparison of Metrics

The point-metric used for comparing classifiers should reflect what is desired. For the rare class problem, we take a stand that it is desired to achieve a high value of recall (R^c) and precision (P^c) for the given rare class C. We now present an analysis of a few different metrics from the perspective of how much weight do they give to the improvements in recall and precision. Specifically, we are looking for a metric that can be improved by improving both recall and precision of the rare class, and not just either one of the two.

We compare seven metrics: Accuracy (Acc), F-measure (F^c), Geometric mean of recall and precision of C ($GMRP^c$), Sum of recall and precision of C ($SORP^c$), Average information score (I_a), Geometric mean of recalls ($GMOR$), and sum of recalls (SOR).

Note that we use $SORP^c$ because it is equivalent to the AUC metric of a single classifier based on the ROC curve plotted as R^c vs. $(1 - P^c)$, just as SOR is equivalent to AUC of the R^c vs. $(1 - R^{nc})$ curve.

Let us assume that α denotes the ratio of number of NC examples to number of C examples in the test set; i.e., $\alpha = n^{nc}/n^c$. Higher the value of $\alpha > 1$, more rare is class C. We first show that each of the metrics can be uniquely expressed in terms of R^c and P^c, and α. Using definition of α, and the definitions of recalls and precisions in Equations 1– 4, it can be verified that

$$R^{nc} = 1 - \frac{R^c}{\alpha}\left(\frac{1}{P^c} - 1\right)$$

Using this expression of R^{nc}, we get following:

$$\begin{aligned}
F^c &= 2\,R^c\,P^c/(R^c + P^c)\\
GMRP^c &= \sqrt{R^c\,P^c}\\
SORP^c &= R^c + P^c\\
GMOR &= \sqrt{R^c\left(1 + \frac{R^c}{\alpha} - \frac{R^c}{\alpha P^c}\right)}\\
SOR &= 1 + R^c\left(1 + \frac{1}{\alpha} - \frac{1}{\alpha P^c}\right)
\end{aligned}$$

For expressing I_a and Acc in terms of R^c, P^c, and α, we need following expressions:

$$\begin{aligned}
n_{00} &= R^c\,n^c = R^c\,n/(\alpha+1)\\
n_{01} &= n^c - n_{00} = (1 - R^c)\,n/(\alpha+1)\\
n_{10} &= n_{00}\left(\frac{1}{P^c} - 1\right) = \frac{R^c\,n}{\alpha+1}\left(\frac{1}{P^c} - 1\right)\\
n_{11} &= n^{nc} - n_{10} = n - \frac{n}{\alpha+1}\left(1 - R^c + \frac{R^c}{P^c}\right)
\end{aligned}$$

Using these expressions and Equation 5, I_a can be written as

$$\begin{aligned}
I_a &= \frac{R^c}{\alpha+1}\left(2 - \frac{1}{P^c}\right)\log(\alpha+1) + \\
&\quad \left(1 - \frac{1}{\alpha+1}\left(2 - 2R^c + \frac{R^c}{P^c}\right)\right)\log\left(\frac{\alpha+1}{\alpha}\right)
\end{aligned}$$

Similarly, Acc can be written as:

$$Acc = \frac{n_{00} + n_{11}}{n} = \frac{1}{\alpha+1}\left(\alpha + 2\,R^c - \frac{R^c}{P^c}\right)$$

Now, we want to compare a classifier that achieves (R_1^c, P_1^c) values on a given data-set with a classifier that achieves (R_2^c, P_2^c) values on the same set. To simplify notation, we drop the superscript c, with an understanding that all the recall and precision values are with respect to the rare class C. For each of the metrics (denoted generically by $\mathcal{M}$) being compared, higher $\mathcal{M}(R, P)$ means better classifier performance. We want to see that given $R_1 > R_2$, what kind of a relationship is required between P_1 and P_2 to ensure that $\mathcal{M}_1 > \mathcal{M}_2$. Note that similar analysis can be done by assuming $P_1 > P_2$, and checking the relationship between R_1 and R_2. We just present the analysis for the former case. Let us use $\beta = R_1/R_2$ ($\beta > 1$).

Using the expressions of each metric in terms of R and P as stated above, some algebraic transformations (a bit involved for some metrics) can be performed to show the following relationships between P_1 and P_2 for each metric:

$$\begin{aligned}
F &: \frac{1}{P_1} < \frac{1}{P_2} + \frac{1}{R_2}\left(1 - \frac{1}{\beta}\right) \\
GMRP &: \frac{1}{P_1} < \frac{\beta}{P_2} \\
SORP &: \frac{1}{P_1} < \frac{1}{R_2\,(1-\beta) + P_2} \\
GMOR &: \frac{1}{P_1} < 1 + \frac{1}{\beta^2}\left(\frac{1}{P_2} - 1\right) + \frac{\alpha}{\beta R_2}\left(1 - \frac{1}{\beta}\right) \\
SOR &: \frac{1}{P_1} < \frac{1}{\beta P_2} + (\alpha + 1)\left(1 - \frac{1}{\beta}\right) \\
Acc &: \frac{1}{P_1} < 2 - \frac{2}{\beta} + \frac{1}{\beta P_2} \\
I_a &: \frac{1}{P_1} < 2 + \frac{1}{\beta P_2} - \frac{2}{\beta}
\end{aligned}$$

Let us denote by L_{P_1} the lower limit on P_1 value, that each of these above expressions is imposing. Interesting point to note is that although the expressions for I_a and Acc involve α, the expressions for L_{P_1} do not depend on α. Moreover, both I_a and Acc impose identical lower limit on P_1.

4 Empirical Comparison of Metrics

In this section, we provide an empirical analysis using the relationships derived in the previous section. Figures 2 and 3 show plots of L_{P_1} for various metrics for different values of R_2 (constant over a column of graphs) and P_2 (constant over a row of graphs). The X-axis of each graph is R_1 and L_{P_1} is on the Y-axis.

The first thing to notice from Figure 2 is that SOR and $GMOR$ depend on the rarity level of C. As the rarity goes up (α increases), these metrics impose very little restriction on P_1, as L_{P_1} dips significantly below the P_2 level. Hence, these are not suitable metrics for judging the rare class performance.

All the other metrics are independent of α. We want to see which of these metrics allows reasonable limits on P_1 for various combinations of R_2 and P_2. Using Figure 3, the following observations can be made about their relative performance.

- $SORP$ is not a good metric, because its performance varies dramatically over various plots. In most cases, except for combinations of higher P_2 values and lower R_2 values ($P_2 = 0.8, R_2 = 0.1$), it imposes one of

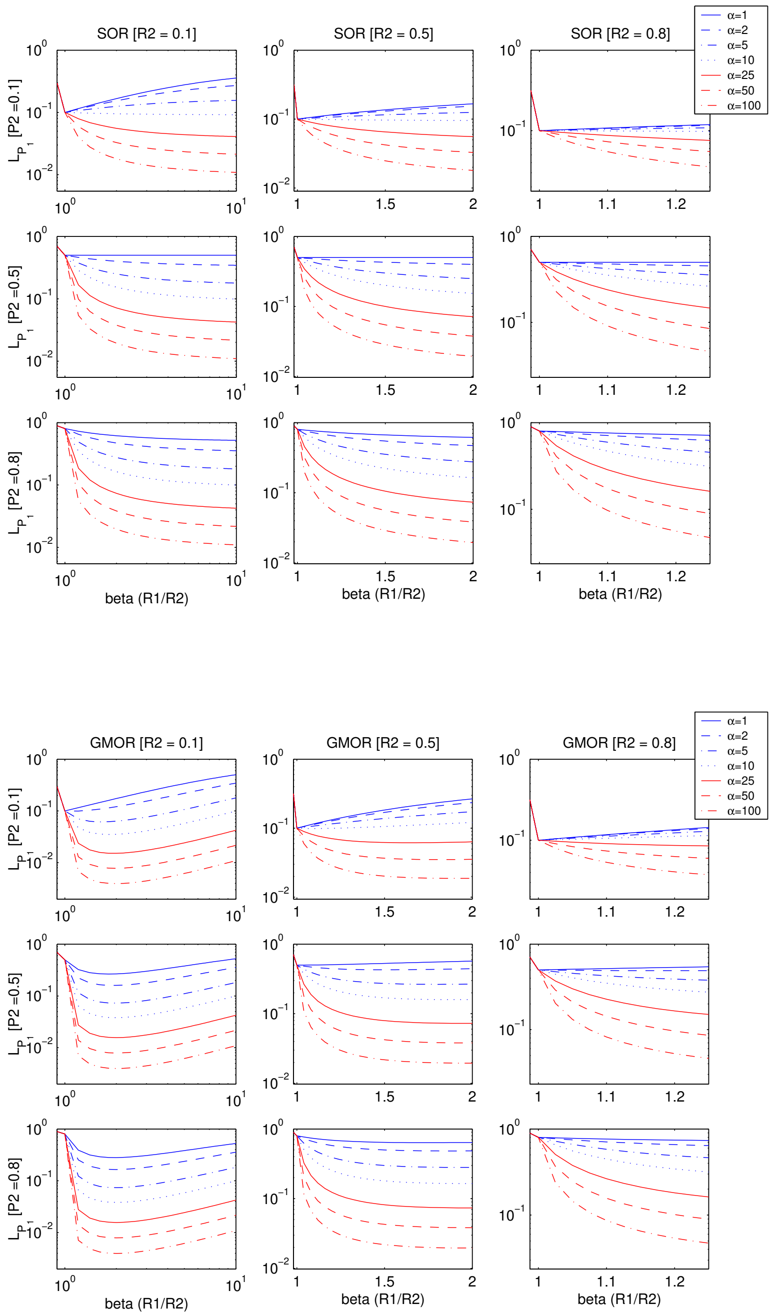

Figure 2. Behavior of Sum-of-Recalls (SOR) and Geometric-Mean-of-Recalls (GMOR) metrics for various combinations of recall and precision performance. Note that SOR is equivalent to AUC of a single classifier for the R^c vs. $(1 - R^{nc})$ ROC curve.

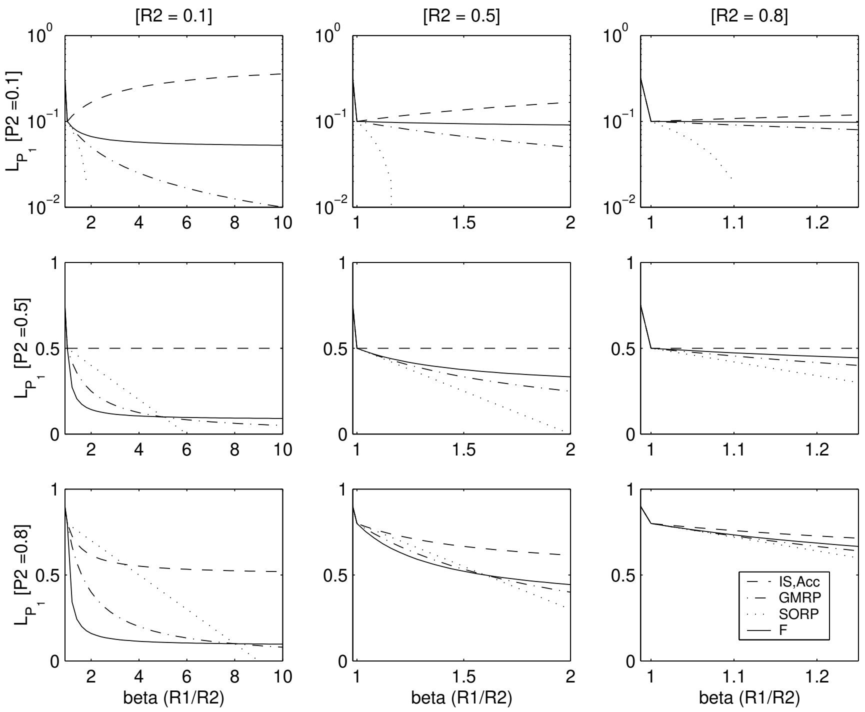

Figure 3. Behavior of various metrics who impose a lower limit on P_1 that is independent of rarity of C. Refer to text for notation. Note that SORP is equivalent to AUC of a single classifier for the R^c vs. $(1 - P^c)$ ROC curve.

the lowest restrictions on P_1.

- For higher values of R_2, most other metrics are similar. F_1 imposes slightly higher limits on P_1 than $GMRP$ does.

- I_a and Acc impose the most stringent requirements on P_2. For lower values of P_2 ($P_2 \leq 0.5$), there is a region in the graph where for any level of improvement in R_1 over R_2, even with $P_1 > P_2$, the metrics can be lower. For example, look at the row of $P_2 = 0.1$. This is not a desired property from rare class perspective, because improving both recall and precision (both significantly according to some statistical test) is an accomplishment when the problem is tougher; but, it is not getting rewarded by I_a and Acc. Another way to look at it is from the graphs in the row of $P_2 = 0.5$. They show that any value of $P_1 > P_2$ will improve the metric, irrespective of the level of improvement in recall. This is again a stringent requirement that tends to favor precision more than recall.

- F_1 achieves a balance between the limits imposed by $GMRP$ and I_a, Acc, for lower values of P_2 and for higher values of R_2. Only for higher values of P_2 and lower values of R_2, does $GMRP$ start getting better in some lower range of β, determined by a threshold that increases as P_2 increases. Here by *better*, we mean the L_{P_1} value for $GMRP$ lies between the F-measure and I_a and Acc. But, for most combinations of R_2 and P_2, F-measure is better than $GMRP$.

The above discussion indicates that ideally one should compare classifiers based on different metrics for different combinations of ranges of the recall and precision levels. But, in practice such a multi-metric comparison strategy is difficult to justify. Two candidates emerge as strong ones, F metric for most of regions in the recall-precision space and $GMRP$ for the low recall (< 0.5), high precision (> 0.5) combinations. F metric has some other advantages: its acceptability in many domains such as text categorization (where rare class problem are encountered very often) and it theoretical derivation in [15] to make it conform to the six conditions of an additive conjoint relational structure [10] imposed on two proportion-based quantities[2]. Later in Section 5, we compare its performance with $GMRP$ (along with other metrics) to observe how they behave on various real-world and synthetic rare class problems.

4.1 Proposed Classifier Comparison Strategy

One drawback of comparing classifiers using just the composite metric such as F-measure is that one cannot directly apply tests of significance to it to determine the confidence level of the comparison. Our primary goal is to achieve improvements in recall as well as precision for the rare class. In order to serve this goal more directly, we now propose a new classifier comparison strategy.

If a classifier achieves significantly better levels for either recall or precision, with respect to some statistical test of significance, then we regard that classifier as a better classifier. We use *p-test* [16] to compare improvements in recall and precision. On a given dataset with n^c examples of C, let classifiers A and B achieve (R_A, P_A, F_A) and (R_B, P_B, F_B) values of recall, precision, and F-measure for C, respectively. Let A predict n_0^A examples as C (sum of true and false positives), and B predict n_0^B examples as C. Given this, the p-test is applied as follows: and P_A with P_B as follows:

$$
\begin{aligned}
Z_R &= \frac{R_A - R_B}{\sqrt{2R(1-R)/n^c}}, \\
Z_P &= \frac{P_A - P_B}{\sqrt{P(1-P)(1/n_0^A + 1/n_0^B)}},
\end{aligned}
$$

where $R = \frac{R_A + R_B}{2}$ and $P = \frac{n_0^A \times P_A + n_0^B \times P_B}{n_0^A + n_0^B}$. Here, we are using the fact that P_A (resp. P_B) is a *proportion* obtained on population n_0^A (resp. n_0^B), and R_A and R_B are proportions on n^c. If $Z_R \geq 1.96$, then we regard

[2]We use the definition of F with respect to the recall (R^c) and precision (P^c) of the rare class, but the general theory of F metric can be applied to define it in terms of R^c and R^{nc} also.

R_A as being significantly better than R_B at the 95% confidence level We denote this relationship by $R_A \gg R_B$. If $Z_R \leq -1.96$, then $R_A \ll R_B$. Similar tests can be applied to compare precision values: if $Z_P \geq 1.96$, then $P_A \gg P_B$ and if $Z_P \leq -1.96$, then $P_A \ll P_B$. On the other hand, if $|Z_R| < 1.96$, then A and B have comparable recall, denoted by $R_A \sim R_B$. If $|Z_P| < 1.96$, then $P_A \sim P_B$.

Note that the confidence interval value of 1.96 should be modified if a confidence level other than 95% is desired. Moreover, this confidence interval is based on the z-test of statistical significance that holds in general if the underlying sample size is ≥ 30. If the values of n^C or $n_0^A + n_0^B$ are < 30, then the tables for t-test should be used to determine the appropriate confidence interval for a given confidence level. The size of the sample size can be used to determine the degrees of freedom required for applying a t-test.

We call classifier A better than B ($A > B$) if either of the following criteria is satisfied:

- $R_A \gg R_B$ and $P_A \sim P_B$; or
- $R_A \gg R_B$ and $P_A \gg P_B$; or
- $R_A \sim R_B$ and $P_A \gg P_B$;

Similarly, one can determine if $B > A$. If both conditions $R_A \sim R_B$ and $P_A \sim P_B$ hold, then $A \sim B$.

It may happen that one metric (R or P) is significantly better and the other (P or R) is significantly worse. Only in such cases, we resort to the composite metric of F-measure. A composite metric does not have any interpretation such as *proportion* or *probability* that either recall and precision have; hence, we cannot apply any statistical significance test to its values. So, some heuristic needs to be applied, such as an improvement in the composite metric's value by at least 1% is required to call a classifier better. The choice of using F-measure for resolving conflicts is supported in the next section, where we present validating arguments that it is the most suitable metric for a wide range of recall, precision, and rarity values.

5 Validation of Analytical Results

In this section, we use the strategy proposed in earlier section to validate the relationships between various metrics that were derived analytically and empirically in Sections 3 and 4.

We use a number of synthetic problems that were designed to test the performance of a classifier on rare class problems. The exact description of the data-sets (denoted by snc) can be found in [8]. For the purposes of this paper, it suffices to know that we choose datasets to have same level of rarity (5% proportion of target class C), but varying levels of learning difficulty as determined by the maximum achievable recall and precision levels.

We form the second set of problems using various datasets at the UCI machine learning repository [1]. In particular we form binary problems for each class with a proportion of $< 35\%$ present in the datasets: thyroid (allhyper, allhypo, ann-thyroid), annealing, chess (king-rook-king), LED, letter recognition, waveform, pima indians diabetes, Boston housing, Statlog (satimage, shuttle, vehicle, segment, DNA), and contraceptive method choice (cmc).

For each problem, we need to apply a classification algorithm to choose the best recall-precision combination according to some composite metric M. Ideally, one should use a classifier that yields a probabilistic score for each decision, and then for each threshold value on the score, we can compute the value of the point metric M. The threshold yielding maximum value of M is chosen. We can also use an algorithm that learns various models on a given problem, and then choose the model that gives the best value of the metric M. We use the latter approach here by using the AdaCost algorithm [8] that uses PNrule [7] as its base learner. For each problem, we run two instances of this algorithm, one that is expected to produce high recall models, and other that is expected to produce high precision models. This is done so as to get a wide range of recall-precision combinations. We refer to the use of each instance of an algorithm to each problem as one *experiment*.

For each experiment and for a given metric M, we choose the model that yields the best value of M on a validation dataset (which is a uniform random sample of the training set) and apply this model to the test-set to record the R_M and P_M values on the test set.

We apply the classifier comparison tests described in Section 4.1 to the classifiers C_{M_1} and C_{M_2} chosen by metrics M_1 and M_2, respectively. There are five possibilities:

- Equal (E): $C_{M_1} \sim C_{M_2}$;
- Worse (W): $C_{M_1} < C_{M_2}$;
- Better (B): $C_{M_1} > C_{M_2}$;
- Recall Worse, Precision Better (P): $R_{M_1} \ll R_{M_2}$, $P_{M_1} \gg P_{M_2}$;
- Recall Better, Precision Worse (R): $R_{M_1} \gg R_{M_2}$, $P_{M_1} \ll P_{M_2}$;

We compare C_F, the classifier chosen by F-measure with C_M for other metrics M, where M can be one of I_a, Acc, SOR, $GMOR$, $SORP$, or $GMRP$. For each experiment, we determine the category of the pair-wise metric comparison as one of E, W, B, P, or R. So, on a particular experiment, if the comparison yields $B = 1$, then we can say that the classifier chosen by F-measure is better than the classifier chosen by M. If on the other hand, we get $W = 1$, then F-measure chooses a worse classifier, and so on for the other categories based on the definitions of E, R, and P above. We add up the values of E, W, B, P, and R over all the experiments and show the aggregated relative performance in the results that we present now.

Table 2 shows the performance over 184 experiments on the UCI datasets and 96 experiments on the synthetic datasets. The very first observation to make is that I_a and Acc are equivalent as far as choosing the best classifier.

Metric	UCI	Synthetic
I_a	165-1-12-0-6	43-5-6-0-42
Acc	165-1-12-0-6	43-5-6-0-42
SOR	144-2-17-21-0	33-2-20-41-0
$GMOR$	144-2-17-21-0	33-2-20-41-0
$SORP$	173-5-2-0-4	43-4-4-21-24
$GMRP$	177-3-0-0-4	74-6-1-6-9

Table 2. Comparing Metrics on Real-world and Synthetic Datasets. Format: E-W-B-P-R.

This is expected according to our analysis. SOR and $GMOR$ are also choosing equivalent models. As Figure 2 indicated, both are dependent primarily on the proportion of the rarer class in the dataset. Although there is a slight variation in the analytical limits they impose on P_1, this variation is negligible with respect to the very low limit values, and the variation between the disappears for moderate levels of R_2 and P_2. This may be the reason that they both choose identical models. Henceforth, we focus on only I_a and SOR as representatives of the two categories.

The second observation is that the R number in the F vs. I_a comparison is higher, indicating that in many cases (6 for UCI and 42 for synthetic) I_a and Acc choose a model which has lower recall and higher precision than the model chosen by F. This is expected from the plots of Figure 3, which show that I_a imposes the most stringent limits on P_1 values, thus forcing a classifier to achieve higher precision. Moreover, the B number is higher than W number is these comparisons, indicating that there are more cases where F chooses an overall better than I_a according to our pure recall and precision based comparison criterion.

The F vs. SOR comparison yields higher number of cases in the P category. So, it chooses a model with higher recall than the model chosen by F. This again is expected from our analysis. The high numbers in B category show that SOR often chooses models worse than F.

The models chosen by $SORP$ show more agreement with F on UCI datasets, but less agreement on the synthetic datasets. It is interesting to note its performance on synthetic datasets, because they have much rarer classes (5% proportion in the population) and have a much wider range of recall and precision levels than the UCI datasets. There is no clear direction provided here. In some experiments, $SORP$ behaves like I_a by emphasizing high precision, while in similar number of experiments, it behaves like SOR by emphasizing high recall. Also, in few cases it yields better model than F and in few cases, a worse model. As our plots of Figure 3 indicated, performance of $SORP$ varies all over the range. Hence, these results further support our argument of Section 4 that

P_F	α	$R_F < 77.3\%$	$R_F \geq 77.3\%$
P_F	$12.5 < \alpha \leq 444.3$	7-0-4-7-0	1-0-2-0-0
< 73.3	$5.4 < \alpha \leq 12.5$	13-0-0-3-0	15-0-2-2-0
	$1.75 < \alpha \leq 5.4$	22-1-0-0-0	13-0-0-0-0
P_F	$1.75 < \alpha \leq 5.4$	8-0-5-1-0	23-1-2-1-0
≥ 73.3	$5.4 < \alpha \leq 12.5$	8-0-2-5-0	11-0-0-1-0
	$12.5 < \alpha \leq 444.3$	5-0-0-1-0	18-0-0-0-0

Table 3. Comparing Effect of Rarity on F vs. SOR Relationship using UCI datasets. Format: E-W-B-P-R.

$SORP$ is not a suitable metric.

Finally, $GMRP$ also exhibits a performance sometimes closer to I_a and sometimes closer to SOR, but the degree of this variation is much smaller as compared to $SORP$. Also, there are more cases falling under the E category, which means that F and $GMRP$ choose equivalent models in most cases. This supports our earlier observation in Section 4 that both $GMRP$ and F-measure are similar for moderate to high levels of recall and precision. Their differences are expected to manifest for lower R_F values, which we will validate shortly.

In the next two subsections, we break down the results of Table 2 into various ranges of rarity and recall-precision levels. to present a clearer picture of these relationships between F-measure and other metrics.

5.1 Effect of Rarity

Table 3 shows the relationship between models chosen by SOR and F-measure for various levels of rarity and recall-precision combinations using the UCI datasets which have a wide variation of both. The range of $\alpha = n^{nc}/n^c$ (defined in Section 3) values present in the UCI datasets was divided into three bins so as to have similar number of experiments in each range. Similarly, the ranges of R_F and P_F values (i.e. the values of models chosen by F-measure) are divided independently to yield two similar-sized bins.

The results for each combination of R_F and P_F indicate that as the proportion of the target class C increases in the datasets (α and rarity decrease), the classifiers chosen by SOR and F begin to be more and more similar (the fraction of cases of type E increases with respect to the other cases). For more rare classes, SOR tends to either choose worse classifiers (cases in B category increase) or it emphasizes more recall (cases in P category increase). This effect is especially pronounced for the lower values of R_F. For higher R_F and P_F values (lower right quadrant of the table), the differences in two metrics become less crucial, even with the variation in rarity. All

P_F	α	$R_F < 77.3\%$	$R_F \geq 77.3\%$
P_F < 73.3	$12.5 < \alpha \leq 444.3$	15-1-1-0-1	3-0-0-0-0
	$5.4 < \alpha \leq 12.5$	11-0-1-0-4	17-0-2-0-0
	$1.75 < \alpha \leq 5.4$	17-0-6-0-0	11-0-2-0-0
P_F ≥ 73.3	$1.75 < \alpha \leq 5.4$	14-0-0-0-0	27-0-0-0-0
	$5.4 < \alpha \leq 12.5$	14-0-0-0-1	12-0-0-0-0
	$12.5 < \alpha \leq 444.3$	6-0-0-0-0	18-0-0-0-0

Table 4. Comparing Effect of Rarity on F vs. I_a Relationship using UCI datasets. Format: E-W-B-P-R.

these observations indicate that when it comes to evaluating classifiers on tougher[3] rare class problems, F metric is a better choice when compared to SOR (and by equivalence $GMOR$).

According to our analysis presented in Sections 3 and 4, the effect of rarity is more crucial for the SOR and $GMOR$ metrics, but it does not affect the relative performance of F and other metrics as much. This is shown for the I_a metric in Table 4. There is very little variation between the relative values of E, W, B, P, and R within each quadrant of the R_F and P_F values. Similar behavior was observed for $SORP$ and $GMRP$ metrics.

5.2 Effect of Recall and Precision Levels

For various levels of recall and precision levels, Tables 5(a)-(d) show the relationship between models chosen by F with models chosen by I_a, SOR, $SORP$, and $GMRP$, respectively. These results are obtained only on the synthetic datasets because we can maintain the same level of rarity while generating datasets with different levels of difficulty based on the achievable recall and precision levels. This makes it easy to study just the effect of recall-precision levels. The ranges of R and P values are independently divided into two similar-sized bins.

First thing to observe is that for high R_F and high P_F combinations, all the metrics except SOR become equivalent to F. But, our primary interest is in tougher problems corresponding to the other three combinations of R_F and P_F.

Table 5(a) indicates that I_a chooses models with low recall and high precision, especially for low R_F values, which indicates that the recall achieved by I_a is significantly lower (lower than R_F) in general. This behavior is observed for lower P_F values also, but that is expected because F-measure is driven by lower of the two values R_F and P_F, hence I_a will yield higher precision models for lower values of P_F. Also I_a has a slight tendency to choose worse models than F for low recall, high precision combinations. But, the most significant result from

[3]By tougher, we mean that achieving both high recall and high precision simultaneously is tough.

P_F	M	$R_F \in [12.1\%, 50.0\%)$	$R_F \in [50.0\%, 93.5\%)$
$[15.7\%, 56.3\%)$	I_a	1-0-2-0-28	8-1-2-0-6
$[56.3\%, 99.8\%)$	I_a	7-3-0-0-7	27-1-2-0-1

(a)

$[15.7\%, 56.3\%)$	SOR	11-1-3-16-0	2-1-2-12-0
$[56.3\%, 99.8\%)$	SOR	7-0-9-1-0	13-0-6-12-0

(b)

$[15.7\%, 56.3\%)$	$SORP$	5-0-2-12-12	7-1-0-8-1
$[56.3\%, 99.8\%)$	$SORP$	4-2-1-0-10	27-1-1-1-1

(c)

$[15.7\%, 56.3\%)$	$GMRP$	24-2-1-1-3	11-1-0-4-1
$[56.3\%, 99.8\%)$	$GMRP$	10-3-0-0-4	29-0-0-1-1

(d)

Table 5. Comparing Effect of Recall-Precision Levels using Synthetic Datasets: (a) F vs. I_a, (b) F vs. SOR, (c) F vs. $SORP$, (d) F vs. $GMRP$. Format: E-W-B-P-R.

this table that makes a case against I_a is the large number (28) for the low recall, low precision combination. This means that for tougher rare class problems, I_a (and by equivalence Acc) will mostly ignore the recall of the rare class. These results corroborate our plots of Figure 3.

Table 5(b) re-affirms the observations made about SOR earlier, that it mostly chooses high recall and low precision models. This table further shows that the SOR behavior is almost independent of the R_F and P_F combinations. As we know from Section 5.1, SOR and $GMOR$ are driven more by the rarity.

$SORP$ was shown to behave erratically in the global results of Table 2. Table 5(c) breaks down that behavior to pinpoint the ranges where it behaves like I_a and/or SOR. For low R_F, high P_F combinations, it chooses models having higher precision and lower recall than models chosen by F. This behavior is similar to the choice of significantly low recall models exhibited by I_a. For high R_F, low P_F combinations, $SORP$ behaves more like SOR by choosing high recall, low precision models, which means it more or less ignores the precision of the model (because its precision is even lower than the low P_F values). For low values of both R_F and P_F, its behavior does not offer any fixed direction. On some experiments, it behaves like I_a and on similar number of experiments it like SOR. All these points act against preferring $SORP$ over F.

Finally, Table 5(d) shows that $GMRP$ is the best competitor to F because in most cases it behaves like F, and in some cases even yields better models than F. But, it has a tendency similar to $SORP$ to a much lesser extent. For low R_F, high P_F combinations, it is similar to I_a on few cases, whereas for high R_F and low P_F values, it is similar to SOR on few cases. This tends to support the behavior shown in Figure 3. Observe that the lower ranges of R_F and P_F in the synthetic datasets is not *very* low. For R_F and P_F values even lower than these, we expect this similarity of $GMRP$ to $SORP$ to become more pronounced.

6 Summary

We address the issue of evaluating the performance of a classifier when it comes to predicting rare classes effectively. We set a goal of rewarding a classifier when it achieves both higher recall and higher precision for the rare class. From this perspective, we have compared a wide range of metrics analytically and empirically, by deriving the dependency of each on the recall and precision of the rare class. Using this analysis and supporting experimental results on various real-world and synthetic datasets, we bring out many valuable insights. First, we demonstrate that the information score (IS) metric derived from the information theory is equivalent to the traditional metric of accuracy, and present results to support that both are not suitable for evaluating the performance of harder problems. In particular, they tend to favor precision, thus choosing models with poorer recall for the rare class. Similarly, the sum-of-recalls (SOR) and geometric-mean-of-recalls ($GMOR$) metrics are observed to be equivalent with strong dependency on the rarity. As the target class becomes more rare, these two tend towards models with significantly lower precision. We demonstrate an erratic behavior of the sum-of-recall-and-precision ($SORP$) metric. It emphasizes recall in some cases and precision in some other cases, thus not providing any clear direction. We argue that the geometric-mean-of-recall-and-precision ($GMRP$) and the F-measure are the most suitable metrics that behave more or less similarly for moderately tough problems. However, $GMRP$ is expected to exhibit a behavior similar to $SORP$ for significantly lower recall and precision values. Hence, we believe that F-measure is the more favorable metric, when it comes to evaluating classifiers for tougher rare class problems. We also propose a new classifier comparison strategy that addresses our primary goal more directly, while bringing in the composite metric of F-measure only to resolve conflicts.

References

[1] C. Blake and C. Merz. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[3] J. P. Egan. *Signal Detection Theory and ROC Analysis*. Series in Cognition and Perception. Academic Press, New York, 1975.

[4] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan. AdaCost: Misclassification cost-sensitive boosting. In *Proc. of Sixth International Conference on Machine Learning (ICML-99)*, Bled, Slovenia, 1999.

[5] C. Hatzis and D. Page. Evaluation metric used for Task 1 of the KDD-CUP'01 competition. In *http://www.cs.wisc.edu/~dpage/kddcup2001*, 2001.

[6] R. C. Holte, N. Japkowicz, C. X. Ling, and S. M. (eds.). Learning from imbalanced data sets (papers from AAAI workshop). Technical Report WS-00-05, AAAI Press, Menlo Park, CA, 2000.

[7] M. V. Joshi, R. C. Agarwal, and V. Kumar. Mining needles in a haystack: Classifying rare classes via two-phase rule induction. In *Proc. of ACM SIGMOD Conference*, pages 91–102, Santa Barbara, CA, 2001.

[8] M. V. Joshi, V. Kumar, and R. C. Agarwal. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In *Proc. of The First IEEE International Conference on Data Mining (ICDM)*, San Jose, CA, Nov 2001.

[9] I. Koronenko and I. Bratko. Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6:67–80, 1991.

[10] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement*, volume 1. Academic Press, London and New York, 1971.

[11] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *Proc. 14th International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.

[12] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[13] F. J. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

[14] J. A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988.

[15] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[16] Y. Yang and X. Liu. A re-examination of text categorization methods. In *22nd Annual International Conference on Information Retrieval (SIGIR'99)*, pages 42–49, Berkley, August 1999.