LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-417778

Class-specific Error Bounds for Ensemble Classifiers

R. Prenger, T. Lemmond, K. Varshney, B. Chen, W. Hanley

October 8, 2009

SIAM Conference on Data Mining (SDM10)
Columbus, OH, United States
April 29, 2010 through May 1, 2010

Disclaimer

Class-specific Error Bounds for Ensemble Classifiers

Ryan J. Prenger[1] Tracy D. Lemmond[1] Kush R. Varshney[2] Barry Y. Chen[1] William G. Hanley[1]

1. Systems and Intelligence Analysis, Lawrence Livermore National Laboratory
2. Laboratory for Information and Decision Systems, Massachusetts Institute of Technology

Abstract

The *generalization error*, or probability of misclassification, of ensemble classifiers has been shown to be bounded above by a function of the mean correlation between the constituent (i.e., *base*) classifiers and their average strength. This bound suggests that increasing the strength and/or decreasing the correlation of an ensemble's base classifiers may yield improved performance under the assumption of equal error costs. However, this and other existing bounds do not directly address application spaces in which error costs are inherently unequal. For applications involving binary classification, Receiver Operating Characteristic (ROC) curves, performance curves that explicitly trade off false alarms and missed detections, are often utilized to support decision making. To address performance optimization in this context, we have developed a lower bound for the entire ROC curve that can be expressed in terms of the class-specific strength and correlation of the base classifiers.

We present empirical analyses demonstrating the efficacy of these bounds in predicting relative classifier performance. In addition, we specify performance regions of the ROC curve that are naturally delineated by the class-specific strengths of the base classifiers and show that each of these regions can be associated with a unique set of guidelines for performance optimization of binary classifiers within unequal error cost regimes.

1 Introduction

Effective classification technologies are vital to systems that learn patterns of behavior from collected data to support prediction and informed decision-making. In particular, human analysts employ classifiers to rapidly sift through millions of samples, identifying those that contain signatures of interest for more in-depth analysis. Many real-world applications that leverage these technologies involve binary (i.e., two-class) classification, in which performance is measured via false alarms (i.e., type I error) and missed detections (i.e., type II error). Generally, the relative costs of these two types of error are inherently unequal, determined a priori by such considerations as limited resources (e.g., time, money, personnel) or actual cost, in terms of loss of capital, loss of life, etc. For example, when combing millions of documents for those relevant to a search query, missed detections may be regarded as an acceptable risk, to avoid overwhelming an analyst with thousands of irrelevant documents (i.e., false alarms). In contrast, when luggage is scanned for the presence of explosives, missed detections would be considered far more costly. Applications such as these are common, and hence, classification methodologies capable of performance optimization within unequal cost regimes are critical.

In this paper, we leverage key elements of Breiman's derivation of a generalization error bound [Breiman2001] to derive novel bounds on false alarms and missed detections. The ultimate objective is to enable the characterization and tuning of factors that affect classifier performance when the error costs are unequal. An analysis of these error-specific bounds leads to a natural partitioning of the ROC curve into three regions, each of which can be associated with a unique set of guidelines for performance optimization. These guidelines will provide insight into ensemble performance within unequal error cost regimes and lead to promising approaches for performance enhancement. Moreover, the bounds will be utilized to establish a lower bound on the entire ROC curve.

In section 2, we will present the three performance regions of the ROC curve along with the bounds on false alarms and missed detections that hold within each region. We will discuss the meaning and implications of our bounds within each performance region and then expand these bounds to the entire ROC curve. In section 3, we will apply Breiman's Random Forest ensemble classifier [Breiman2001] to both the SPECTF and Parkinson's datasets and show that the ROC curve lower bound predicts (1) the shape and trend of the true ROC curve and (2) the relative performance of competing ensembles. Conclusions are presented in Section 4.

2 Class-Specific Error Bounds

The concept of combining multiple models, the cornerstone of ensemble methods, originated as early as 1977 with the combination of two linear regression models by Tukey [Tukey1977, Rokach2009]. With the advent of more sophisticated computer technologies, however, ensemble methodologies have evolved to leverage potentially thousands of base classifiers that are usually instantiations of the same underlying model (e.g., neural networks, decision trees).

An ensemble makes class predictions by propagating a test sample through each base classifier, which assigns a class label, or *vote*, to the sample. Typically, the sample is then assigned to the class receiving the majority vote. However, in cost-sensitive applications, we can threshold the resulting vote

frequencies to enable a classification decision that is sensitive to differing error requirements. Bounding the errors associated with those decisions is of great interest, and is discussed in detail in the following sections.

2.1 Generalization Error

Since their inception, ensemble methodologies have proven to be highly successful at reducing the generalization error in classification [Rokach2009]. Placing a bound on the generalization error is beneficial both for characterizing the performance of ensemble classifiers in the field as well as in motivating efforts at ensemble classifier optimization. Generalization error bounds have previously been derived for ensemble classification methods by [Breiman2001, Garg2002, and Koltchinskii2003].

The bound derived by Breiman is of particular interest, because it incorporates ensemble characteristics that are highly interpretable and may enable the selective tuning of ensemble performance. Specifically, he demonstrated that as the number of base classifiers in the ensemble increases, the generalization error, E, converges and is bounded as follows:

$$E \le \frac{\bar{\rho}\left(1-s^2\right)}{s^2}, \tag{1}$$

where $\bar{\rho}$ denotes the mean correlation of base classifier predictions, and s represents the average strength of the base classifiers[1]. From (1), it is immediately apparent that the bound on generalization error decreases as the base classifiers become stronger and/or less correlated. However, note that (1) does not explicitly characterize the impact of the strength and correlation of base classifiers on class-specific error rates. To this end, we have developed extensions to Breiman's bound that directly address these error rates.

2.2 ROC Curve Performance Regions

As discussed in Section 2.1, vote frequencies generated by the ensemble are used to classify a data sample. When the positive and negative classes are associated with the labels 1 and 0 respectively[2], these votes can be combined to compute a numerical *score*, given by

$$score(\mathbf{x}) = \frac{2}{K}\sum_{k=1}^{K} h_k(\mathbf{x}) - 1, \tag{2}$$

where K equals the number of base classifiers in the ensemble and $h_k(\mathbf{x})$ is the label given by k^{th} base classifier to the input vector $\mathbf{x}$. The score lies within the interval $[-1,1]$ and relates directly to the *margin function*.

Given a collection of votes generated by an ensemble, the margin function measures the degree to which the votes for the correct class exceed the votes for the incorrect class; in essence, it is a measure of confidence. Breiman [Breiman2001] has shown that, for the two-class case, the margin function for an ensemble classifier can be expressed as

$$mg(\mathbf{x},y) = \frac{2}{K}\sum_{k=1}^{K} I\left(h_k(\mathbf{x}) = y\right) - 1, \tag{3}$$

where $I(\cdot)$ is an indicator function and y is the true class label associated with data sample $\mathbf{x}$. Note that for class 1 samples, the score is equal to the margin, and it can be easily shown that for class 0 samples, the score is the negative of the margin.

The scores computed for each class form distributions that can be used to generate a ROC curve. Each point on the ROC curve indicates the false alarm and detection rates of the ensemble classifier, given a fixed decision threshold. Consequently, the curve can be generated by sweeping a decision threshold across the two class-specific score distributions simultaneously, as illustrated in Figure 1. Note that the probability mass to the right of this threshold for the positive and negative class score distributions corresponds to the detection and false alarm rates, respectively.

Breiman [Breiman2001] defines the average strength of the base classifiers as the expected value of the margin function. Leveraging the relationship between the score distributions and the margin function, we can estimate the class-specific strengths, s_0 and s_1, by

$$s_0 = -\mu_0 \text{ and } s_1 = \mu_1, \tag{4}$$

where μ_i is the mean of the score distribution for class i. The overall strength, s, can be written in terms of the class-specific strengths as:

$$s = \frac{n_1 s_1 + n_0 s_0}{n_1 + n_0}, \tag{5}$$

where n_i is the number of class i samples. Thus, the overall strength is a weighted average of the class-specific strengths, and it measures the degree of separation between the means of the score distributions.

The variance σ^2 of the margin function is also related to the strength and correlation of the base

[1] For the sake of brevity, we will often refer to average strength and mean correlation as simply *strength* and *correlation*, respectively.

[2] Under this assumption, we will frequently refer to the positive and negative classes as class 1 and 0, respectively.

classifiers, and can be expressed in general by the following inequality:

$$\sigma^2 \le \bar{\rho}\left(1 - s^2\right). \tag{6}$$

We can write Eq. (6) in terms of the positive and negative classes as follows:

$$\begin{aligned} {\sigma_0}^2 &\le \bar{\rho}_0\left(1 - {s_0}^2\right) \\ {\sigma_1}^2 &\le \bar{\rho}_1\left(1 - {s_1}^2\right) \end{aligned}, \tag{7}$$

where σ_i^2 is the variance of the class i score distribution and $\bar{\rho}_i$ denotes the mean correlation between the base classifiers calculated for the class i samples. Eq. (7) clearly shows that for fixed class-specific strength, reducing (or increasing) the class-specific correlation between the base classifiers can yield a corresponding shift in the variance of the margin function (and hence, the variance of the score distribution) for that class. We will discuss this relationship further in Section 2.3.

2.3 Bounding the ROC Curve

The generalization error bound derived by Breiman regards all errors as equally important, and the decision threshold is implicitly fixed at zero. Hence, this represents a bound on a single point on the ROC curve. To extend this bound to the entire ROC curve, thus bounding performance across all error cost regimes, every decision threshold value must be considered. The one-tailed Chebychev inequality, shown in (8), enables us to derive bounds on the false alarm rate and detection rate in terms of the class-specific strengths and correlations for a given threshold, t.

$$P\left(Z - \mu \ge k\right) \le \frac{1}{1 + \dfrac{k^2}{\sigma^2}}, \text{ for } k > 0 \tag{8}$$

For example, a bound on the false alarm rate (FAR) for a decision threshold $t \in [\mu_0, 1]$ can be derived (see Appendix 1 for a complete derivation) from (8) via the variable substitution $t = \mu_0 + k$, and is given by

$$FAR = P\left(Z_0 \ge t\right) \le \frac{1}{1 + \dfrac{\left(t - \mu_0\right)^2}{{\sigma_0}^2}}, \; t \in [\mu_0, 1]. \tag{9}$$

From the relationships given in (4) and (7), equation (9) can be expressed in terms of the strength and mean correlation for the negative class and is given by

$$FAR = P\left(Z_0 \ge t\right) \le \frac{1}{1 + \dfrac{\left(t + s_0\right)^2}{\bar{\rho}_0\left(1 - {s_0}^2\right)}}, \; t \in [-s_0, 1]. \tag{10}$$

Similar derivations can be performed for both tails of Chebychev's inequality, yielding an upper or lower bound for both the false alarm and detection rates over different subintervals of $[-1, 1]$. These subintervals naturally partition the class-specific score distributions (and hence, the ROC curve) into three distinct regions

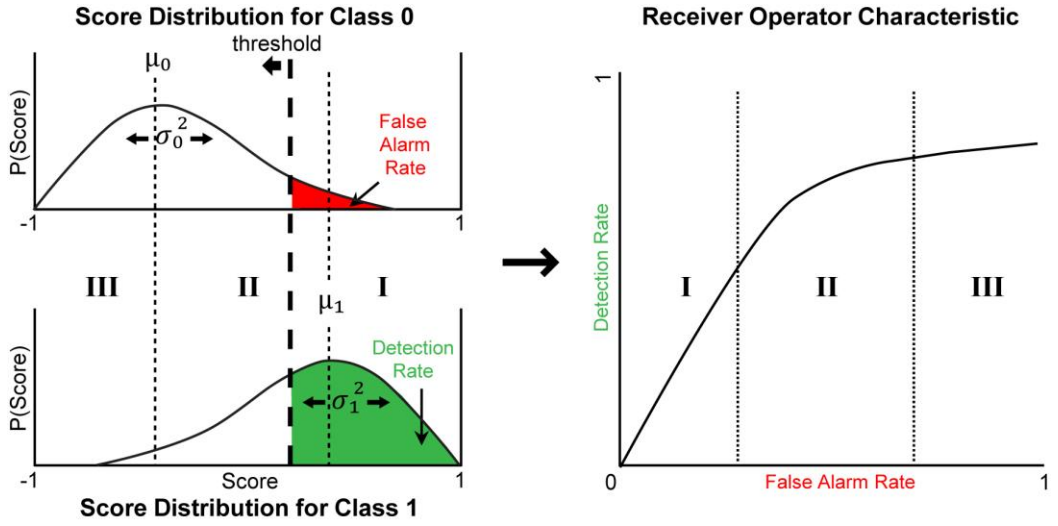

Figure 1. Performance Regions of a ROC Curve
Score distributions for both classes are shown (left), indicating the three performance regions delineated by the class-specific score means. The ROC curve generated by sweeping a decision threshold across the score distributions is also shown (right), with the corresponding performance regions indicated.

Table 1. Bounds on the false alarm rate (FAR) and detection rate (DET) for each of three performance regions on the ROC curve.

Region III $t \in [-1, -s_0]$	**Region II** $t \in [-s_0, s_1]$	**Region I** $t \in [s_1, 1]$
$FAR \geq \dfrac{1}{1+\dfrac{\bar{\rho}_0\left(1-s_0^2\right)}{\left(t+s_0\right)^2}}$	$FAR \leq \dfrac{1}{1+\dfrac{\left(t+s_0\right)^2}{\bar{\rho}_0\left(1-s_0^2\right)}}$	$FAR \leq \dfrac{1}{1+\dfrac{\left(t+s_0\right)^2}{\bar{\rho}_0\left(1-s_0^2\right)}}$
$DET \geq \dfrac{1}{1+\dfrac{\bar{\rho}_1\left(1-s_1^2\right)}{\left(t-s_1\right)^2}}$	$DET \geq \dfrac{1}{1+\dfrac{\bar{\rho}_1\left(1-s_1^2\right)}{\left(t-s_1\right)^2}}$	$DET \leq \dfrac{1}{1+\dfrac{\left(t-s_1\right)^2}{\bar{\rho}_1\left(1-s_1^2\right)}}$

that are delineated by the means μ_0 and μ_1. For threshold values within each region, denoted I, II, and III as shown in Figure 1, the corresponding false alarm and detection rate bounds characterize ensemble performance in terms of the class-specific strength and the correlation associated with the base classifiers. These bounds are presented in Table 1.

Careful inspection of the bounds presented in Table 1 reveals the desired characteristics of the class-specific strength and correlation of the base classifiers that will yield bounds most favorable to ensemble performance. For example, in Region I, when strength is held fixed, it is clear that decreasing the correlation for the negative class samples, $\bar{\rho}_0$, decreases the upper bound for the false alarm rate, potentially resulting in improved performance. Similarly, *increasing* the correlation for the positive class samples, $\bar{\rho}_1$, will increase the upper bound of the detection rate. Though improved performance is not guaranteed, these bounds suggest guidelines for tuning the ensemble to produce more favorable conditions for minimizing class-specific errors.

It should be noted here that the region-specific guidelines derived from these bounds are highly consistent with the intuition gleaned from the score distribution diagram in Figure 1. As we observed in Section 2.2, for a fixed strength, an increase in the class-specific correlation can lead to an increase in the variance of the corresponding score distribution.

Figure 1 illustrates that when the decision threshold is very high in Region I, an increase in the spread (i.e., variance) of the class 1 score distribution, for a fixed mean, may increase the number of scores lying to the right of threshold, thus increasing the detection rate. This is a form of stochastic resonance, in which adding variability to the system improves performance. Intuitive arguments similar to that above can be made regarding the bounds in the remaining regions. Note that in all cases when correlation is held fixed, higher strength for both classes produces a greater separation of the score means and may yield improved performance. For a fixed strength, the guidelines for tuning class-specific correlation inferred from the bounds for each region are summarized in Table 2.

Table 2. Tuning class-specific correlation.

Region	**Guidelines**
I	$\bar{\rho}_0 \downarrow$ and $\bar{\rho}_1 \uparrow$
II[3]	$\bar{\rho}_0 \downarrow$ and $\bar{\rho}_1 \downarrow$
III	$\bar{\rho}_0 \uparrow$ and $\bar{\rho}_1 \downarrow$

$$DET = \left[1 + E_{MISS}\left(1 - \sqrt{\frac{E_{FAR}\left(1-FAR\right)}{FAR}}\right)^{-2}\right]^{-1}$$
$$E_{FAR} = \frac{\bar{\rho}_0\left(1-s_0^2\right)}{\left(s_1+s_0\right)^2}, \quad E_{MISS} = \frac{\bar{\rho}_1\left(1-s_1^2\right)}{\left(s_1+s_0\right)^2}, \qquad (11)$$
$$\text{for } FAR \in \left[\frac{E_{FAR}}{E_{FAR}+1}, 1\right]$$

We can use the error bounds derived for Region II to compute a lower bound for the entire ROC curve; a complete explanation is included in Appendix 2. This lower bound can be expressed as shown in (11).

E_{FAR} and E_{MISS} are the key components of the derived ROC lower bound. Figure 2 illustrates the effects of reducing the quantities E_{FAR} and E_{MISS} on the lower bound of the ROC curve. Specifically, at sufficiently low false alarm rates, the ROC lower bound can only be improved by decreasing E_{FAR}. Similarly, for sufficiently high false alarm rates, the lower bound can only be improved by decreasing E_{MISS}.

[3] Note that the guidelines for Region II are similar to those derived from Breiman's bound, where the decision threshold is implicitly fixed at zero.

2.4 Leveraging Class-Specific Error Tradeoffs

Because Regions I and III correspond to low false alarm and missed detection rates, respectively, they are of great interest for the many real world applications that involve extreme differences in error cost. Like Breiman's bound, the error bounds derived for these regions are relatively loose; hence, they serve most effectively as an intuitive guide to performance optimization.

As shown in Table 2, the error bounds for Regions I and III yield *opposing* guidelines with respect to class-specific mean correlation. Specifically, if the class-specific correlations could be effectively controlled for fixed means, performance within these regions of the true ROC curve could be explicitly traded off based upon relative error costs..

In addition, the ROC lower bound may also be influenced via manipulation of the class-specific strength and correlation, as evidenced by Eq. (11). Near the boundaries of Region II, decreasing the quantities E_{FAR} and E_{MISS} yields a shift in the ROC lower bound, as illustrated in Figure 2. Interestingly, the correlation for class 1 samples plays no role in E_{FAR}, while the correlation for class 0 samples plays no role in E_{MISS}. Thus, Eq. (11) suggests that to shift the bound near the boundary between Regions I and II, we must balance the strength and correlation for class 0 and increase the strength for class 1 as much as possible, without regard to the class 1 correlation. A similar argument holds for shifting the bound near the boundaries of Regions II and III.

It is important to realize that when the positive and negative classes are sufficiently well separated, as shown in Figure 3, the entire ROC curve may reside in Region II, where high strength and low correlation for both classes result in lower error bounds. Any attempts to increase class-specific correlation under these conditions would prove counterproductive.

Achieving a sufficient degree of control to enable the correlation of the base classifiers to be tuned for each class, as proposed above, presents a significant challenge in general. However, in Section 3, we will investigate an approach to increasing the correlation over both classes that applies specifically to the Random Forest, and we will examine its impact on the three performance regions.

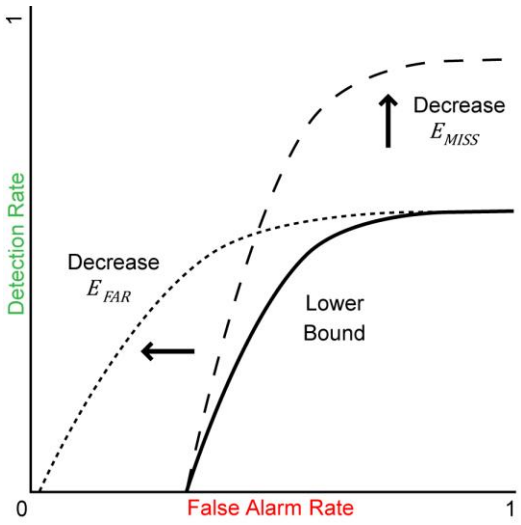

Figure 2. The effects of decreasing E_{FAR} and E_{MISS} on the ROC Lower Bound are shown.

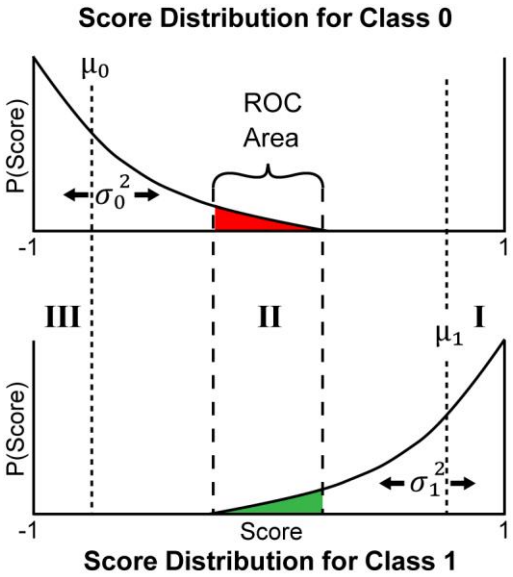

Figure 3. Class-specific score distributions where the ROC curve lies entirely in Region II are shown. The only range of thresholds used to draw the ROC curve lie within the overlap between the distributions.

3. Empirical Analysis using Random Forests

The Random Forest (RF) is an ensemble methodology that utilizes decision trees as its base classifiers [Breiman2001]. A decision tree is constructed via a series of hierarchical univariate node decisions. Prior to training an RF, a split dimension, *m*, is specified which determines the number of features considered at each node. It can be shown empirically that higher split dimensionality results in stronger and more correlated trees.

In the following experiments, we utilized two publicly available data sets to demonstrate (1) the effects of varying the split dimension on class-specific strength and correlation, and (2) the ability of the ROC lower bound to predict the relative performance of competing forests (i.e., the degree to which the true ROC curves and their corresponding bounds are similarly "nested"). Each experiment consisted of 101 trials, initiated using different random seeds, to enable

an assessment of statistical significance. Within each trial, Random Forests were trained using split dimensions ranging from 1 to 15. Each forest in these studies was composed of 500 Gini-based decision trees [Breiman1984], and the resulting ROC curves were computed using the *out-of-bag* samples.

The data sets of interest in these studies come from the UCI Machine Learning Repository [Asuncion 2007]. The first of these is the SPECTF Heart data set, consisting of features from cardiac Single Proton Emission Computed Tomography (SPECT) images used to classify patients as normal or abnormal. This data set consists of 267 samples and 44 features. We also present results for the Parkinson's data set, which consists of 197 samples and 23 features (speech signals). The task in this case was to determine whether a subject has Parkinson's Disease.

3.1 The Similarity Metric

To quantify the degree to which the true ROC curves and their bounds are similarly "nested", we developed a similarity measure computed as described below.

For each of 10,000 evenly distributed values of the decision threshold, *t*, we computed the correlation coefficient between false alarm and detection rates and their corresponding bounds across the different Random Forest split dimensions. The average of these correlation coefficients was used as a measure of similarity between the ROC curves and bounds at a particular *t*. The overall similarity measure was obtained by averaging the pointwise similarity measures over all values of *t*.

To determine whether the resulting similarity was statistically significant, a *p*-value was computed from the 101 trials. We used $p < 0.05$ to determine significance.

3.2 Predictive Capability of the ROC Bound

The ROC curves and corresponding bounds obtained by applying the Random Forest to the SPECTF data set are shown in Figure 4. Note that all three performance regions contain some portion of the ROC curves, and it can be visually observed within each region that the "nesting" of the ROC curves is consistent for both the true curves and their bounds. Moreover, the trend and shape of the bounds strongly resemble the true ROC curves. From a quantitative standpoint, the median similarity measure was 0.6676, and was found to be statistically significant ($p < 0.01$). The ROC bounds correctly predicted that lower split dimensions would yield better performance in Regions I and II, while higher split dimensions would be favored in Region III.

We can interpret these results further by examining the class-specific strengths and correlations, as well as E_{FAR} and E_{MISS}, as a function of split dimension, shown in Figure 5. With regard to E_{FAR} and E_{MISS}, recall that lower values reflect increased performance at the boundaries between Regions I and II, and II and III respectively. In this case, the lower split dimensions produced lower values for E_{FAR} and E_{MISS}. This is consistent with Figure 4, which shows that lower split dimensions are superior throughout all of Region II, including areas near the boundaries.

These results also demonstrate that an increase in the correlation between the base classifiers can benefit the ensemble in Region III. On this data set, increasing the split dimension has a negligible effect on the strength, but it increases the correlation of the base classifiers for both classes.

In Region III, the ROC bounds produced for higher split dimensions are slightly favored over those for lower dimensions. It is shown in Appendix 3 that, for very low thresholds, the relative behavior of the ROC bounds in Region III is increasingly determined by the ratio E_{MISS}/E_{FAR}, rather than by E_{MISS} alone. Specifically, lower E_{MISS}/E_{FAR} values tend to be associated with better performance in Region III. The same is true of E_{FAR}/E_{MISS} for very high thresholds in Region I. These ratios are also plotted as a function of split dimension in Figure 5. We have observed that when both class-specific correlations are increased by roughly the same amount, under the assumption of fixed strength, the ratio E_{MISS}/E_{FAR} will decrease. This provides strong evidence that the increase in the class-specific correlations played a role in improving the bound in Region III.

The same experiments that were performed for the SPECTF data were performed for the Parkinson's data set, and the resulting ROC curves and bounds are shown in Figure 6. The bounds clearly predict the variability observed in the true ROC curves due to the split dimension. The median similarity value, 0.5004, was statistically significant ($p < 0.01$). Note that in this example, no portion of the true ROC curves is present in Region I. Additionally, because only a few points on the ROC curve lie in Region III, the performance of forests in the low false alarm rate region will not be substantially improved by increasing the base classifier correlation for class 1 alone.

To provide further insight, the class-specific strengths and correlations, E_{FAR}, E_{MISS}, and their ratios are plotted as a function of split dimension in Figure 7. It is clear that increasing the split dimension increases

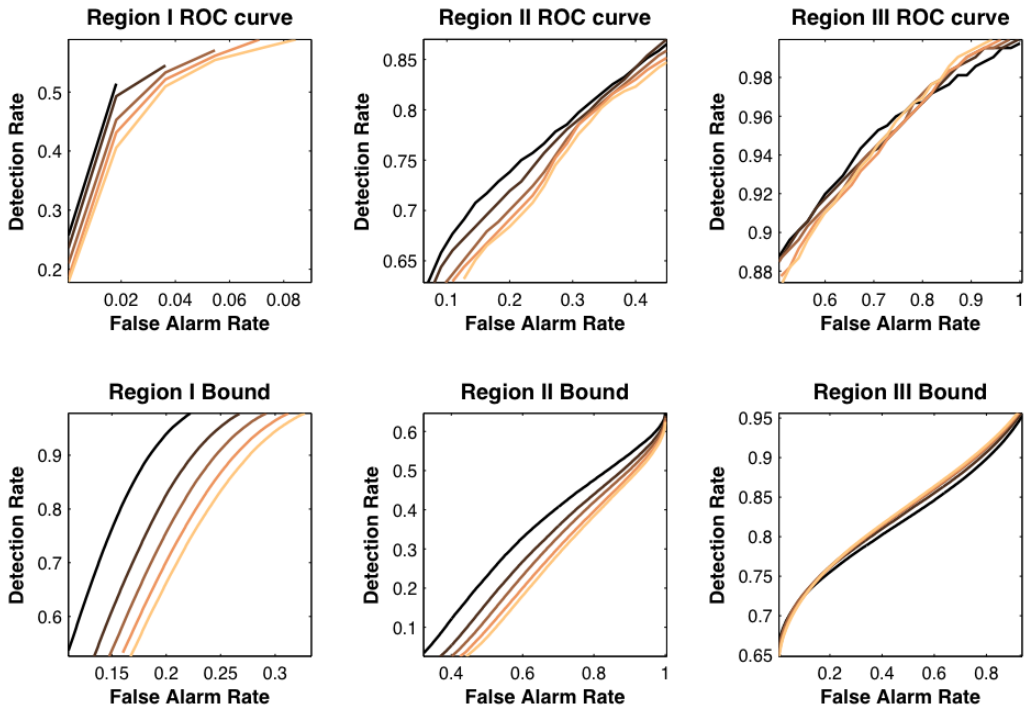

Figure 4. ROC bounds predict relative RF performance across split dimensions on the SPECTF Data Set. ROC curves for five Random Forests trained using different split dimensions in the range [1,15] (lighter color implies higher dimension) are plotted for the three performance regions.

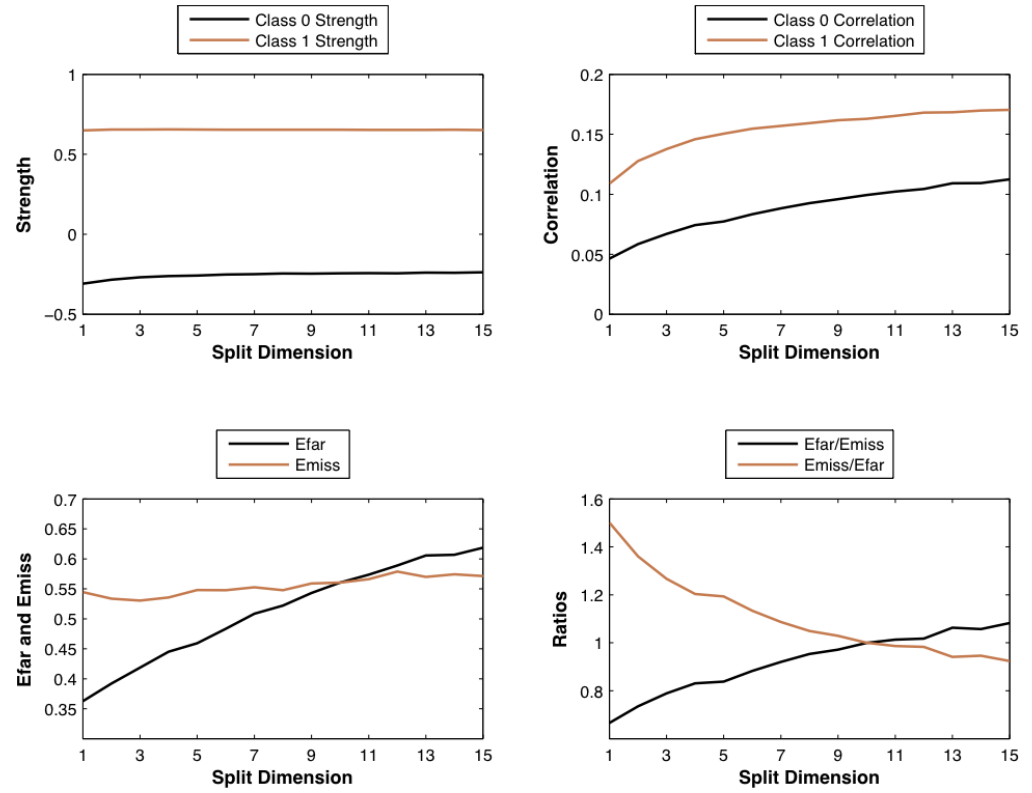

Figure 5. Strength, Correlation, E_{FAR}, and E_{MISS} on the SPECTF Data Set.

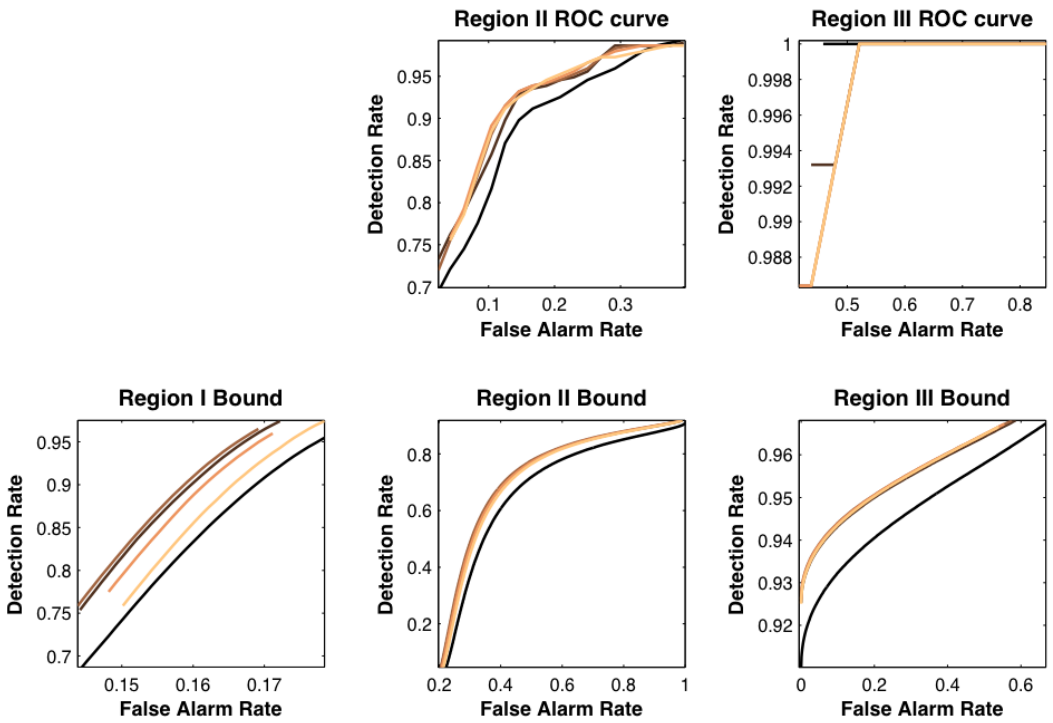

Figure 6. ROC bounds predict relative RF performance across split dimensions on the Parkinson's Data Set. ROC curves for Random Forests trained using different split dimensions in the range [1, 15] (lighter color implies higher dimension) are plotted for each of the three performance regions.

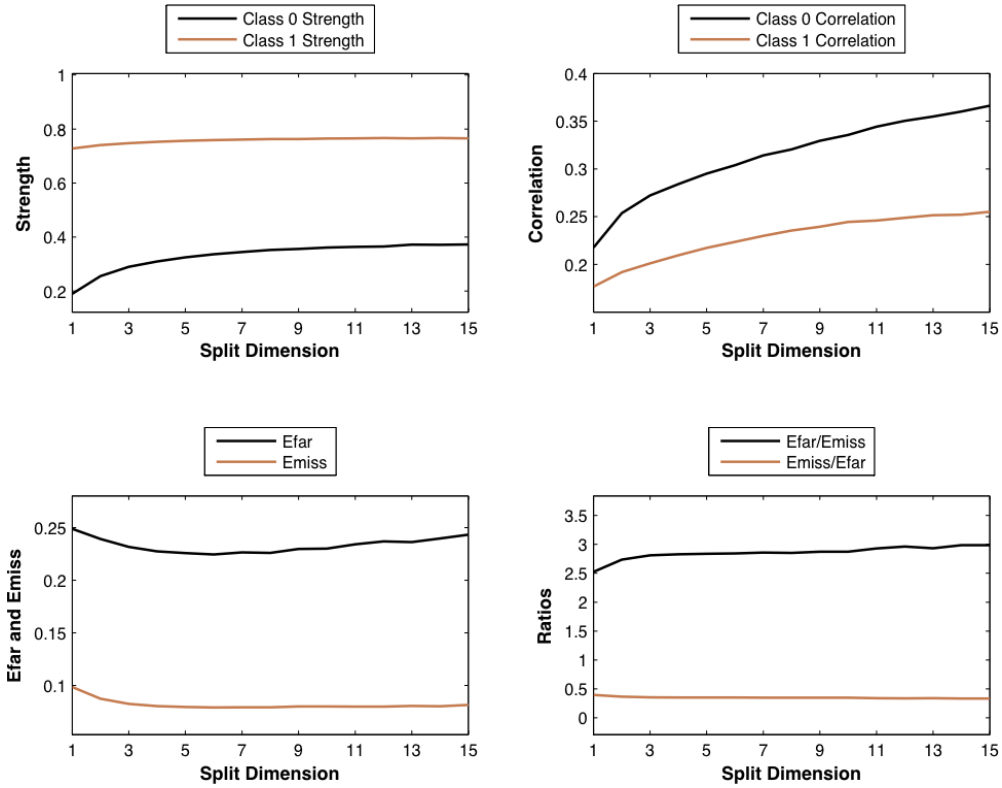

Figure 7. Strength, Correlation, E_{FAR}, and E_{MISS} on the Parkinson's Data Set.

the strengths and correlations for both classes. Interestingly, both E_{FAR} and E_{MISS} are minimized at intermediate values of split dimension. This was found to be consistent with Figure 6, which indicates that the bounds for the intermediate split dimensions are superior across Region II.

In this case, because there is so great a separation between the score distributions, there are few threshold values high or low enough to generate bounds in Regions I and III. The bounds that can be drawn are very near the boundary with Region II, and behave much like those within Region II.

4. Conclusions

To address classification performance optimization for real-world applications that have unequal error costs, we have extended Breiman's generalization error bound to the entire ROC curve. Our analysis has shown that there are distinct regions of the ROC curve (Region I - the extreme low false alarm rate region and Region III - the extreme low miss rate region) in which different class-specific correlations are desired. Specifically, for decision thresholds lying in each of these regions, increasing the correlation on a specific class may improve the performance of the ensemble classifier. However, not surprisingly, there is a clear trade off in performance optimization between Regions I and III, because they respond to opposing guidelines with respect to increasing or decreasing the class-specific correlations. Thus, when the strengths and correlations are altered in a class-specific way, there should be a strong motivation to optimize performance in exactly one of these regions.

A comparison of Random Forests, trained using different split dimensions, has shown that the ROC lower bounds are predictive of relative classifier performance. Specifically, comparing the ROC curves and corresponding ROC bounds generated from these different RFs, we demonstrated that the bounds are predictive of the actual ROC curves for all three performance regions of interest.

This research suggests a number of methods for improving the performance of classifiers within the defined regions on the ROC curves. When the ensemble provides poor separation between distributions (and hence Regions I and III are defined), we expect techniques such as asymmetric boosting to provide a performance trade-off between these regions. Additionally, Breiman originally suggested bagging as a method of decreasing the correlation between base classifiers while preserving their strength. Our analyses suggest that resampling techniques that favor sampling of one class over the other could provide a mechanism for tuning class-specific correlation.

Note, however, that these ultra-low error rate regions in which increased class-specific correlations are desirable will disappear as the performance of the classifier improves (i.e., score distributions become more separated). Hence, methods that trade off performance in Region I in favor of performance in Region III will be ineffective (and possibly counterproductive) in situations where performance is already quite high. On the other hand, if the entire ROC curve lies in Region II, far from the boundaries of Regions I and III, improving performance on either class via tuning of strength or correlation will likely result in an improvement on the other class as well.

For others in the community, this research may prove helpful for the investigation and evaluation of other ensemble techniques that attempt to optimize performance in a class-specific way (such as [Fan1999] and [Masnadi-Shirazi2007]).

Appendix 1. Bounds on False Alarm Rates and Detection Rates

When constructing the ROC curve, the False Alarm Rate (FAR) is the probability that a score exceeds some threshold t from the class 0 empirical score distribution. Similarly the Detection Rate (DET) is the probability that a score exceeds t from the class 1 empirical score distribution. These rates can be expressed as:

$$FAR = P(Z_0 \geq t) \text{ and } DET = P(Z_1 \geq t), \qquad \text{(A1)}$$

where Z_0 and Z_1 are random variables representing the class-specific scores for a particular sample.

We can place bounds on these quantities using the one-tailed Chebychev inequality:

$$P(Z - \mu \geq k) \leq \frac{1}{1 + \frac{k^2}{\sigma^2}} \quad \text{for} \quad k > 0, \qquad \text{(A2)}$$

for some k, where Z has mean μ and finite variance σ^2. Eq. (A2) states that values of Z are not likely to be much greater than the mean. Eq. (A2) can be transformed to a statement about the probability of a value being larger than a threshold t via the variable substitution, $t = k + \mu$, which yields the following inequality:

$$P(Z \geq t) \leq \frac{1}{1 + \frac{(t-\mu)^2}{\sigma^2}} \quad \text{for} \quad t > \mu. \qquad \text{(A3)}$$

Note that Eq. (A3) only applies to the tail of the score distribution where $t > \mu$. The other tail of this distribution similarly gives us a bound on $P(Z \leq t)$, and

we can subtract both sides of the inequality from 1 to yield an inequality describing the region $t < \mu$:

$$P(Z \geq t) \geq \frac{1}{1+\frac{\sigma^2}{(t-\mu)^2}} \quad \text{for} \quad t < \mu. \tag{A4}$$

Equations (A3) and (A4) now give us two limits on the probability that a random variable *Z* will be greater than the threshold *t* in terms of the mean and variance of the distribution.

Now if we take Z to be the class-specific scores of an ensemble classifier, the variance can be related to the correlation between the base classifiers. [Breiman2001] showed that the variance of scores is related to the correlation between base classifiers and their strength, as follows:

$$\sigma^2 \leq \bar{\rho}(1-\mu^2), \tag{A5}$$

where $\bar{\rho}$ is Breiman's measure of mean correlation between the base classifiers in the ensemble. The expressions on the right hand side of inequalities (A3) and (A4) are monotonically increasing and decreasing functions of the variance, respectively (assuming nonzero variance). Hence, we can substitute the right hand side of (A5) in place of the variance in (A3) and (A4) without violating the inequalities. The resulting bounds, in terms of the mean correlation of base classifiers, are given by

$$P(Z \geq t) \leq \frac{1}{1+\frac{(t-\mu)^2}{\bar{\rho}(1-\mu^2)}} \quad \text{for} \quad t > \mu, \tag{A6}$$

$$P(Z \geq t) \geq \frac{1}{1+\frac{\bar{\rho}(1-\mu^2)}{(t-\mu)^2}} \quad \text{for} \quad t < \mu. \tag{A7}$$

As discussed in the main body of the text, Eq. (A8) relates the class-specific strength to the score distributions as follows:

$$s_0 = -\mu_0 \text{ and } s_1 = \mu_1. \tag{A8}$$

Hence, we can use the expressions given by (A6 – A8) to bound the False Alarm Rate and Detection Rates in (A1) in terms of the class-specific strengths and correlations. This resulting bounds are summarized in Table 1.

Appendix 2. ROC Lower Bound

At each value of the threshold *t* in Region II, the detection and false alarm rates must satisfy the expressions in Table 1. Although the value of *t* is constrained to be between $-s_0$ and s_1, it is easily shown that the Region II bound on detection rate goes to 0 as *t* goes to s_1, and the corresponding bound on the false alarm rate goes to 1 as *t* goes to $-s_0$. Hence, these bounds can be used to generate a lower bound for the entire ROC curve by plotting the bound values at every threshold. We can derive an equation for this curve by first setting the Region II inequalities in Table 1 to equalities, then solving for *t*. The resulting system of equations can be solved to obtain the smallest possible value of DET as a function of FAR, given by

$$DET = \left[1+\frac{\bar{\rho}_1(1-s_1^2)}{\bar{\rho}_0(1-s_0^2)}\left(\frac{(s_1+s_0)}{\sqrt{\bar{\rho}_0(1-s_0^2)}}-\sqrt{\frac{1-FAR}{FAR}}\right)^{-2}\right]^{-1} \tag{A9}$$

and the constraint $-s_0 < t < s_1$ can be used to obtain constraints on the values of FAR and DET. Specifically, if we substitute $-s_0$ and s_1 for *t* into the Region II FAR bound found in Table 1, we obtain:

$$\frac{1}{1+\frac{(s_1+s_0)^2}{\bar{\rho}_0(1-s_0^2)}} < FAR < 1. \tag{A10}$$

Constraints on DET can be obtained in a similar fashion. Finally, by substituting the values E_{FAR} and E_{MISS}, given below, into (A9) and (A10), we arrive at Eq. (11) found in the main text.

$$E_{FAR} = \frac{\bar{\rho}_0(1-s_0^2)}{(s_1+s_0)^2} \text{ and } E_{MISS} = \frac{\bar{\rho}_1(1-s_1^2)}{(s_1+s_0)^2} \tag{A11}$$

Appendix 3. E_{FAR} and E_{MISS} ratios in Regions I and III

The same derivations performed in Appendix 2 can be performed using the FAR and DET bounds for Regions I and III found in Table 1, and an expression for DET as a function of FAR can be derived. However the resulting expressions will not be upper or lower bounds on the entire ROC curve in these cases. For example, in Region I, the inequalities produced by this derivation are upper bounds on both DET and FAR. They describe the best possible DET at the worst possible FAR, given the class-specific strengths and mean correlations. However, this function does provide

information about the behavior of the ROC curve in Region I.

For Region I, the function for the highest DET and highest FAR, given the class-specific strengths and mean correlations, are given by:

$$DET = \left[1 + \frac{1}{E_{MISS}} \left(\sqrt{\frac{E_{FAR}(1-FAR)}{FAR}} - 1 \right)^2 \right]^{-1}$$

$$\text{for } FAR \in \left[\frac{E_{F\min}}{E_{F\min}+1}, \frac{E_{FAR}}{E_{FAR}+1} \right], \text{ where} \qquad \text{(A12)}$$

$$E_{F\min} = \frac{\bar{\rho}_0 (1 - s_0^2)}{(1+s_0)^2}$$

The key point to notice about equation (A12) is that as the FAR approaches zero, the equation for DET approaches

$$DET = \frac{E_{MISS}}{E_{FAR}} \frac{FAR}{(1-FAR)}. \qquad \text{(A13)}$$

Hence, as the false alarm rate approaches zero (i.e., the extreme boundary of Region I), the detection rate is increasingly determined by the ratio of E_{FAR} to E_{MISS}. Specifically, lower ratios of E_{FAR}/E_{MISS} may produce better detection rates. However, this only holds for false alarm rates on the interval specified in (A12), so equation (A13) is merely suggestive of the behavior of the ROC curve when the threshold is very high (far from the boundary between regions I and II).

Similar arguments apply to Region III, where we instead can calculate the function for the lowest DET and lowest FAR, given the class-specific strengths and correlations. The result suggests that far from the boundaries between Regions II and III, a lower ratio of E_{MISS}/E_{FAR} is preferable.

Acknowledgements

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

A. Asuncion, and D.J. Newman, (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Chapman and Hall, 1984.

L. Breiman, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

W. Fan, S. Stolfo, J. Zhang, and P. Chan. "Adacost: Misclassication Cost-sensitive Boosting." In *ICML*, 1999.

A. Garg, V. Pavlovic, and T. S. Huang, "Bayesian Networks as Ensemble of Classifiers", *16th International Conference on Pattern Recognition (ICPR'02)*, vol. 2, pp.779-784, 2002.

T. K. Ho, "Random Decision Forest", in *Proc. Of the 3rd International Conference on Document Analysis and Recognition*, pp. 278-282, 1995.

V. Koltchinskii, D. Panchenko, and F. Lozano, "Bounding the Generalization Error of Convex Combinations of Classifiers: Balancing the Dimensionality and the Margins", Annals of Applied Probability, vol. 13, no. 1, pp. 213-252, 2003.

H. Masnadi-Shirazi and N. Vasconcelos. "Asymmetric Boosting". In *ICML*, 2007.

T. D. Lemmond, A. O. Hatch, B. Y. Chen, D. A. Knapp, L. J. Hiller, M. J. Mugge, and W. G. Hanley, "Discriminant Random Forests," *Proceedings of the 2008 International Conference on Data Mining (DMIN'08)*, July 2008.

L. Rokach, "Taxonomy for Characterizing Ensemble Methods in Classification Tasks: A Review and Annotated Bibliography", in *Computational Statistics and Data Analysis*, 53, pp.4046-4072, 2009.

J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.