Practical Ensemble Classification Error Bounds for Different Operating Points

Kush R. Varshney, *Member*, *IEEE*, Ryan J. Prenger,
Tracy L. Marlatt, Barry Y. Chen, and William G. Hanley

Abstract—Classification algorithms used to support the decisions of human analysts are often used in settings in which zero-one loss is not the appropriate indication of performance. The zero-one loss corresponds to the operating point with equal costs for false alarms and missed detections, and no option for the classifier to leave uncertain test samples unlabeled. A generalization bound for ensemble classification at the standard operating point has been developed based on two interpretable properties of the ensemble: strength and correlation, using the Chebyshev inequality. Such generalization bounds for other operating points have not been developed previously and are developed in this paper. Significantly, the bounds are empirically shown to have much practical utility in determining optimal parameters for classification with a reject option, classification for ultralow probability of false alarm, and classification for ultralow probability of missed detection. Counter to the usual guideline of large strength and small correlation in the ensemble, different guidelines are recommended by the derived bounds in the ultralow false alarm and missed detection probability regimes.

Index Terms—Cantelli inequality, random forests, receiver operating characteristic, reject option

1 Introduction

Ensemble classification methods based on bagging, boosting, arcing, and bacing have become first choice algorithms for numerous signal analysts and data mining practitioners [1], [2], [3], [4]. Automatic classification algorithms can quickly sift through large amounts of data that a human would otherwise have to do. For example, consider an analyst at an information technology (IT) outsourcing provider who must label service tickets into categories such as *network*, *operating system*, or *hardware* to analyze what problems are frequently occurring at client installations [5]. The load on the analyst can be reduced if an automatic classifier can accurately label many tickets. As another example, consider an analyst at a national security agency who must find instances of terrorist chatter in petabytes of internet traffic to then be investigated.

In such scenarios, the classifier is used to support the human analyst. Thus, classification requirements are different than they would be for a classifier operating by itself and induce objectives besides zero-one loss. Classification with a reject option, classification for ultralow probability of false alarm, and classification for ultralow probability of missed detection are all relevant in decision-support scenarios [6], [7]. If the automatic classifier can classify half the IT service tickets with high accuracy, rejecting the tickets on which it is uncertain and leaving them for the analyst to manually classify, the workload of the human is halved. An automatic classifier that identifies chatter instances containing terrorist conversations with modest false alarm probability would overwhelm investigators who need to focus their attention on definite instances; a classifier with ultralow false alarm probability is thus appropriate. There are other applications in which ultralow missed detection probability is appropriate, such as in detecting dirty bombs. In either case, the tradeoff between the probabilities of false alarm and missed detection is important to take into account.

A generalization bound for ensemble classification reported in [8] is a function of the *strength* and the *correlation* of the ensemble, which are statistics of its margin distribution. The margin distribution has been used in developing generalization bounds for ensemble classifiers in other work as well [9], [10]. The strength and correlation parameters have semantic meaning about the base classifiers: individually accurate and collectively diverse. This generalization bound on probability of classification error (zero-one loss) is loose, but practically it does show a strong correlation with testing error on many real-world data sets. However, it is only applicable to probability of classification error and does not apply to probability of rejection, probability of false alarm, or probability of missed detection.

As our primary contribution in this paper, we examine the different modes of classifier operation and develop generalization bounds for these modes. Specifically, we develop bounds for reject option risk and for the receiver operating characteristic (ROC) that are functions of strength and correlation. The bounds we develop in this paper are based on the Cantelli inequality; a key feature of the bounds is their practical significance. Another contribution of our work is that we show that on real-world data sets, the

- K.R. Varshney is with the Business Analytics and Mathematical Sciences Department, IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598. E-mail: krvarshn@us.ibm.com.
- R.J. Prenger, T.L. Marlatt, and B.Y. Chen are with the Systems and Intelligence Analysis Section, National Security Engineering Division, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550. E-mail: {prenger1, marlatt1, chen52}@llnl.gov.
- W.G. Hanley is with Exponent, Menlo Park, CA 94025.
E-mail: wghanley3@hotmail.com.

Manuscript received 3 Mar. 2012; revised 14 Sept. 2012; accepted 19 Oct. 2012; published online 8 Nov. 2012.
Recommended for acceptance by H. Wang.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2012-03-0139.
Digital Object Identifier no. 10.1109/TKDE.2012.219.

1041-4347/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

structure of the bound functions mirrors the structure of the empirical reject option risk and ROC functions in practically important ways that allow us to use the bounds to set the rejection threshold or determine which ensemble to use for low false alarm or missed detection probability. A key finding in this paper based on the bounds we develop is that in the ultralow missed detection and false alarm regimes, the guidelines on strength and correlation for optimizing performance are different than those for plain misclassification error. Preliminary versions of this work are reported in [11], [12].

Our contribution, practical generalization bounds on probability of rejection, false alarm, and missed detection for ensemble classification schemes such as bagging and random forests have not, to the best of our knowledge, been reported previously. Generalization bounds for these criteria in the statistical learning theory literature are not well suited to analyzing ensembles and are not practical. Practical bounds for ensembles do not exist for the criteria that arise at different operating points. Recent work on ensemble classification with imbalanced data shares some similarity to our work. A more detailed comparison to the literature and discussion of the novelty of our work is presented in the next section.

The remainder of the paper is organized as follows: In Section 2, we further describe related work and contextualize our contribution. In Section 3, we describe the setup of the ensemble classification problem, including classification with a reject option and classification at different false alarm/missed detection operating points. In Section 4, we develop and interpret bounds for the reject option risk and for the receiver operating characteristic. In Section 5, we present empirical results that show the value of the bounds from a practical perspective. In Section 6, we conclude.

2 Related Work

As mentioned in Section 1, ensemble classification methods have become popular approaches among practitioners and have consequently seen several enhancements and variations recently. Some examples reported in these Transactions include orthogonalizing ensembles of decision trees via Fourier analysis [13], dealing with drifting distributions of data [14], [15], and including a clustering step prior to ensemble classification [16].

The main topic of the work herein is generalization bounds. Generalization bounds within the statistical learning theory paradigm of empirical processes (e.g., Vapnik-Chervonenkis bounds and Rademacher bounds) have been derived for classification with a reject option, ROC analysis, and related nonzero-one loss classification [17], [18], [19], [20], [21], [22]. However, these bounds are not particularly suitable for ensemble classifiers. Additionally, it is well known that generalization bounds derived within the paradigm of empirical processes are not practically useful as expressions to be optimized [23].

A stylized averaging classifier with a reject option is presented in [24], but mainly to allow for empirical process-based proofs. The reject option enters more as a proof technique than anything else, and a generalization error bound for the reject option risk is not given. It is written that it “might be possible to adapt the theory presented in this paper to give a rigorous analysis for the performance of bagging and other ensemble methods” but we are not aware of such an adaptation thereafter.

The consistency of ensemble classifiers is investigated in [25], but the theoretical analysis therein is not for operating points other than the standard one. Similarly, studies of generalization for ensemble classifiers have found that small classification error occurs when the ensemble has high diversity (i.e., low correlation) among the constituent base classifiers and high individual strength for the base classifiers [8], [26], [27], [28]; however, these studies have only examined the standard operating point as well. Moreover, they have found that it is not possible to simultaneously have very high diversity and very high individual strength. Note that it is shown later in this paper that at ultralow false alarm and ultralow missed detection operating points, high (class-specific) diversity is not the guideline like it is at the standard operating point.

Thus, overall, we have not seen any related work that develops practical generalization bounds for ensemble classification with a reject option or when either false alarms or missed detections are very costly. There are bounds for nonzero-one criteria from the empirical process paradigm that are not meant for ensemble classification such as bagging and random forests, and are not practically useful. Also, there are practical generalization bounds and characterizations based on strength and diversity of ensembles, but they are not developed for ROC analysis, ultralow false alarm and missed detection regimes, or reject option risk. This missing piece of the literature is where we contribute with this work.

In research on ensemble classification with imbalanced data carried out concurrently with ours that we became aware of after the submission of this paper, it is found that there are regimes in which diversity has an opposing effect on error in the minority and majority classes [29]; the six situations in [29, Table 3] share many similarities with the three regions summarized in Table 1 in Section 4.3. The similarity is not surprising because costs and priors both appear in analogous roles in the Bayes-optimal operating point or threshold of binary hypothesis testing [7]. Either false alarms or missed detections having very high costs is analogous to either negative or positive examples having very low prior probability.

3 Problem Setup

In this section, we first give the setup and notation for ensemble classification in general. Then, we describe the setup for classification with a reject option and classification at different operating points of the false alarm and missed detection tradeoff. Finally, we describe the combination of different operating points and a reject option.

3.1 Ensemble Classification

We consider the general supervised binary classification problem in which class labels $y \in \{-1, +1\}$ are to be predicted using feature vectors $\mathbf{x} \in \mathcal{X}$. The classification function $\hat{y}(\cdot) : \mathcal{X} \to \{-1, +1\}$ is learned from labeled training data drawn from a fixed distribution and applied to new, unseen and unlabeled test vectors from the same

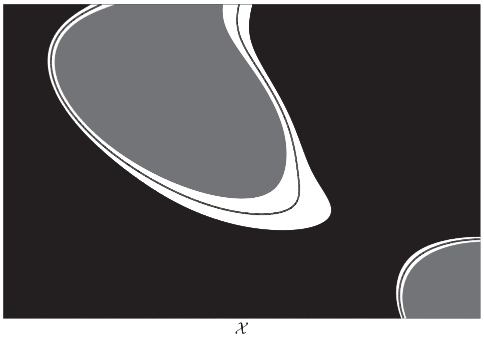


Fig. 1. Illustration of decision regions in feature space. The region $\hat{y} = +1$ is black, the region $\hat{y} = -1$ is gray, the region $\hat{y} =$ reject is white, and the boundary $\phi = 0$ is the black line.

distribution. Each training and test sample is statistically independent from the other samples.

In the specific case of ensemble classification, $\hat{y}$ is composed of base classifiers $\hat{y}_i(\cdot) : \mathcal{X} \to \{-1, +1\}$, $i = 1, \ldots, m$. The overall decision is based on the average classification of the base classifiers. Let the average classification of the base classifiers be the score $\phi \in [-1, +1]$:

$$\phi(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \hat{y}_i(\mathbf{x}). \tag{1}$$

In the usual case, if the score is negative, then the overall decision is $\hat{y} = -1$, and if it is positive, then $\hat{y} = +1$, i.e., $\hat{y}(\mathbf{x}) = \mathrm{sign}(\phi(\mathbf{x}))$. Equivalently, we classify by thresholding the score at zero:

$$\hat{y}(\mathbf{x}) = \begin{cases} -1, & \phi(\mathbf{x}) \leq 0 \\ +1, & \phi(\mathbf{x}) > 0. \end{cases} \tag{2}$$

This classification rule can also be interpreted as majority vote.

Let us define the margin to be $z = y\phi \in [-1, +1]$. Due to the special encoding $y \in \{-1, +1\}$, the margin is negative for incorrect classifications and positive for correct classifications.

3.2 Classification with a Reject Option

Classifications are most uncertain near the boundary between the two classes, which occurs at $\phi(\mathbf{x}) = 0$. In classification with a reject option, rejections are declared in the most uncertain regions of the feature space $\mathcal{X}$, where the score $\phi(\mathbf{x})$ is close to zero. The ensemble classification rule with a reject option is

$$\hat{y}(\mathbf{x}) = \begin{cases} -1, & \phi(\mathbf{x}) \leq -t \\ \text{reject}, & -t < \phi(\mathbf{x}) < t \\ +1, & \phi(\mathbf{x}) \geq t, \end{cases} \tag{3}$$

where $t \geq 0$ is a rejection threshold. In essence, the rejection threshold provides a guard band or padding around the decision regions for $+1$ and -1 with the amount of padding controlled by t, as illustrated in Fig. 1.

The margin is in the range $[-t, +t]$ for rejections. Thinking of the margin as a random variable induced by the random variables $\mathbf{x}$, y, and the learned classifiers $\hat{y}_i$, the probability density function of the margin is denoted $f_z(z)$. Thus, the probability of error

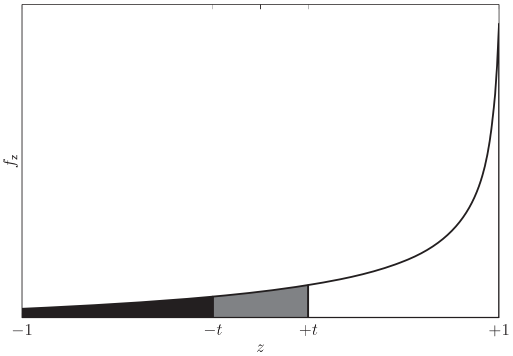


Fig. 2. Illustration of margin distribution $f_z(z)$ marked with rejection threshold t. The area of the black region is the error probability $P_E(t)$. The area of the gray region is the rejection probability $P_R(t)$.

$$\begin{aligned} P_E(t) &= \Pr[z \leq -t] \\ &= \int_{-1}^{-t} f_z(z)dz \end{aligned} \tag{4}$$

and the probability of rejection

$$\begin{aligned} P_R(t) &= \Pr[-t < z < t] \\ &= \int_{-t}^{t} f_z(z)dz. \end{aligned} \tag{5}$$

The error probability $P_E(t)$ is the area of the black region and the rejection probability $P_R(t)$ is the area of the gray region in Fig. 2, an illustration of a margin distribution.

As discussed in [19], a useful measure of performance is the reject option risk:

$$\begin{aligned} L_c(t) &= P_E(t) + cP_R(t) \\ &= \Pr[z \leq -t] + c\Pr[-t < z < t], \end{aligned} \tag{6}$$

where the cost of misclassification is 1, the cost of correct classification is 0, and the cost of rejection is $0 \leq c \leq 1/2$. This reject option risk $L_c(t)$ should be small for good performance.

3.3 Classification at Different Operating Points

Classification errors come in two varieties: false alarms and missed detections. Many classification problems have an alarm state such as the presence of a bomb or terrorist activity, and a nonalarm state in its absence. We associate the class label $+1$ with the alarm state and -1 with the nonalarm state. With this association, a false alarm is when the true label is -1 but the classifier outputs $+1$, and a missed detection is when the true label is $+1$ but the classifier outputs -1. (A detection is when the true label is $+1$ and the classifier also outputs $+1$.) Often, one of the two types of error is more costly than the other, and it is of interest to examine their probabilities separately rather than lumped together in an overall probability of error.

Moreover, in situations of unequal costs, it is imperative to allow the threshold of the ensemble classification rule to deviate from zero, cf. (2). Different thresholds trade missed detection probability and false alarm probability, implemented as

$$\hat{y}(\mathbf{x}) = \begin{cases} -1, & \phi(\mathbf{x}) \leq t \\ +1, & \phi(\mathbf{x}) > t, \end{cases} \tag{7}$$

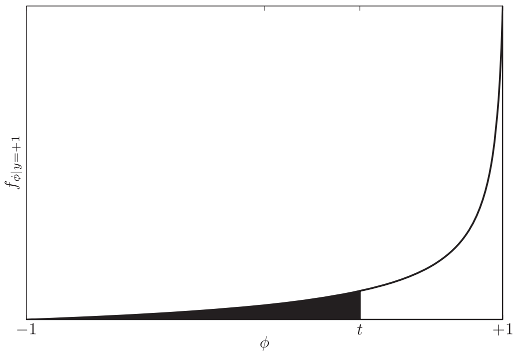


Fig. 3. Illustration of conditional score distribution $f_{\phi|y}(\phi|y=+1)$ marked with threshold t. The area of the black region is the missed detection probability $P_M(t)$. The area of the white region is the detection probability $P_D(t)$.

where $t \in [-1, +1]$ is the classification threshold. Classification performance, i.e., the missed detection and false alarm probabilities, is a function of the conditional distribution of the score ϕ given the true class label y. Illustrations of these conditional distributions are shown in Figs. 3 and 4.

The missed detection probability is

$$\begin{aligned} P_M(t) &= \Pr[\phi \le t \mid y = +1] \\ &= \int_{-1}^{t} f_{\phi|y}(\phi \mid y = +1) d\phi, \end{aligned} \tag{8}$$

and the false alarm probability is

$$\begin{aligned} P_F(t) &= \Pr[\phi > t \mid y = -1] \\ &= \int_{t}^{1} f_{\phi|y}(\phi \mid y = -1) d\phi. \end{aligned} \tag{9}$$

The detection probability is

$$\begin{aligned} P_D(t) &= \Pr[\phi > t \mid y = +1] \\ &= \int_{t}^{1} f_{\phi|y}(\phi \mid y = +1) d\phi. \end{aligned} \tag{10}$$

The threshold t should be set closer to $+1$ for costlier false alarms and closer to -1 for costlier missed detections. The ROC is the parameterized curve $(P_D(t), P_F(t))$ obtained by varying the threshold t from -1 to $+1$.

3.4 Classification at Different Operating Points with a Reject Option

It is seen less often in the literature, but we can also consider the combination of a reject option and different operating points [30]. In this case, there are two thresholds t_1 and t_2, generally not centered around zero. Without loss of generality, let $t_1 \le t_2$. The classification rule for this case is

$$\hat{y}(\mathbf{x}) = \begin{cases} -1, & \phi(\mathbf{x}) \le t_1 \\ \text{reject}, & t_1 < \phi(\mathbf{x}) < t_2 \\ +1, & \phi(\mathbf{x}) \ge t_2. \end{cases} \tag{11}$$

If $t_1 = -t_2$, then we are back to classification with a reject option discussed in Section 3.2. If $t_1 = t_2$, then we are back to classification with different costs for false alarms and missed detections discussed in Section 3.3.

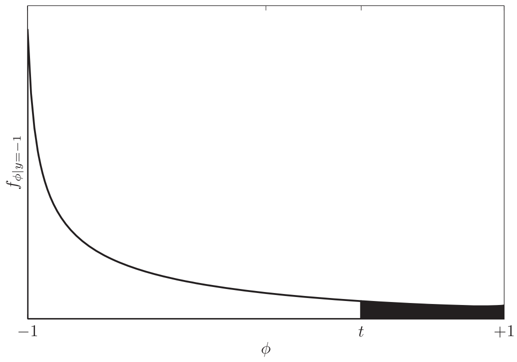


Fig. 4. Illustration of conditional score distribution $f_{\phi|y}(\phi|y=-1)$ marked with threshold t. The area of the black region is the false alarm probability $P_F(t)$.

The probabilities to characterize performance are the probability of missed detection, false alarm, and rejection. Similar to before, they are

$$P_M(t) = \Pr[\phi \le t_1 \mid y = +1] \tag{12}$$

$$P_F(t) = \Pr[\phi \ge t_2 \mid y = -1] \tag{13}$$

$$\begin{aligned} P_R(t) &= \Pr[t_1 < \phi < t_2] \\ &= \Pr[y = +1] \Pr[t_1 < \phi < t_2 \mid y = +1] \\ &\quad + \Pr[y = -1] \Pr[t_1 < \phi < t_2 \mid y = -1]. \end{aligned} \tag{14}$$

4 Bounds Based on Strength and Correlation

In this section, we first provide definitions of the ensemble properties strength and correlation. Then, we derive and discuss bounds on the reject option risk, probability of missed detection, probability of false alarm, and ROC that are a function of these ensemble properties using the Cantelli inequality. The rationale for using the Cantelli inequality is twofold: First, it makes no distributional assumptions on the data or the margins of constituent base classifiers, and second, it uses only two parameters, distributional mean and variance, leading to understandable bounds based on two parameters, strength and correlation.

4.1 Strength and Correlation

Recall that the ensemble classifier $\hat{y}(\mathbf{x})$ is constructed from base classifiers $\hat{y}_i(\mathbf{x}), i = 1, \dots, m$. The correlation between two base classifiers $i \ne j$ is averaged across all pairs of base classifiers to yield the correlation:

$$\bar{\rho} = \frac{2}{m(m-1)} \sum_{i \ne j} \mathrm{E}[\hat{y}_i(\mathbf{x}) \hat{y}_j(\mathbf{x})]. \tag{15}$$

It indicates how diverse the base classifiers are. The strength is the expected value of the margin random variable z:

$$s = \mathrm{E}[z]. \tag{16}$$

It indicates the quality of the individual base classifiers. A relationship between the variance of the margin, the correlation, and the strength of the ensemble shown in [8] is

$$\mathrm{var}(z) \leq \bar{\rho}(1 - s^2). \tag{17}$$

Assuming that $s > 0$, the requirement that base classifiers not be worse than random, the following bound on generalization error is derived in [8] using the Chebyshev inequality and (17):

$$\Pr[y \neq \hat{y}(\mathbf{x})] \leq \frac{\bar{\rho}(1 - s^2)}{s^2}. \tag{18}$$

For small generalization error, correlation should be small and strength should be large. In practice, correlation and strength cannot be optimized in that way. If strength is large, then correlation is also large. For example, if the strength is perfectly 1, the correlation must necessarily be 1. The bound is generally loose, but is often strongly correlated to classification error empirically.

4.2 Bound for Reject Option Risk

We derive a bound for the reject option risk $L_c(t)$ involving the strength and correlation of the ensemble. It may be observed in Fig. 2 that

$$L_c(t) = (1 - c)P_E(t) + c\Pr[z < t]. \tag{19}$$

We bound the probability of error term first, followed by the $\Pr[z < t]$ term; we then combine the two.

The term $P_E = \Pr[z \leq -t]$ is to be bounded using strength and correlation. We use the Cantelli (one-sided Chebyshev) inequality toward this end [31]:

$$\Pr[z - \mathrm{E}[z] \leq -k] \leq \frac{1}{1 + \frac{k^2}{\mathrm{var}(z)}}, \quad k > 0. \tag{20}$$

Letting $k = \mathrm{E}[z] + t$,

$$\Pr[z \leq -t] \leq \frac{1}{1 + \frac{(\mathrm{E}[z]+t)^2}{\mathrm{var}(z)}}, \quad \mathrm{E}[z] > -t, \tag{21}$$

and due to (17) and the definition of s,

$$P_E(t) \leq \frac{1}{1 + \frac{(s+t)^2}{\bar{\rho}(1-s^2)}}, \quad s > -t. \tag{22}$$

With base classifiers having accuracy greater than random guessing ($s > 0$), s must be greater than $-t$ and this constraint need not be further considered.

Now, turning to the second half of the reject option risk expression (19), again by the Cantelli inequality,

$$\Pr[z < t] \leq \frac{1}{1 + \frac{(\mathrm{E}[z]-t)^2}{\mathrm{var}(z)}}, \quad \mathrm{E}[z] > t \tag{23}$$

and also

$$\Pr[z < t] \leq \frac{1}{1 + \frac{(s-t)^2}{\bar{\rho}(1-s^2)}}, \quad s > t. \tag{24}$$

We find a bound for the reject option risk by combining (22) and (24):

$$L_c(t) \leq \frac{1 - c}{1 + \frac{(s+t)^2}{\bar{\rho}(1-s^2)}} + \frac{c}{1 + \frac{(s-t)^2}{\bar{\rho}(1-s^2)}}, \quad s > t. \tag{25}$$

This bound is applicable when the rejection threshold is set below the strength of the ensemble. A threshold value greater than the strength would mean that the classifier is rejecting signals that are "easy" to classify and is not the regime in which the reject option is typically employed. Results in Section 5 show that this bound, although not tight in difference, is quite predictive of the risk behaviors empirically exhibited by ensemble classification with a reject option.

With the goal of small reject option risk, the bound expression (25) may be examined to determine good values for the threshold, strength, and correlation. With fixed rejection cost, strength, and correlation, it is straightforward to determine a closed-form expression for the optimal threshold value $t \in [0, s)$ that minimizes the reject option risk bound. It is a large polynomial expression. Thus, the bound provides a way to set the rejection threshold, the main free parameter in classification with a reject option.

Another analysis that may be considered is to determine guidelines for the strength and correlation of the ensemble with fixed rejection cost and threshold. In this analysis, the idea is to move probability mass from the black area in Fig. 2 to the gray area, or ideally into the white area. The derivative of the $L_c(t)$ bound with respect to s is always negative, so the guideline is to have strength as high as possible. The derivative of the $L_c(t)$ bound with respect to $\bar{\rho}$ is always positive, so the guideline is to have correlation as low as possible. The guidelines of large strength and small correlation are the same as for plain ensemble classification without reject option; further guidelines specific to the reject option may be revealed by higher order analysis [32].

4.3 Bound for Receiver Operating Characteristic

We derive bounds for the detection probability, the false alarm probability, and the ROC in this section. The derivations follow the same pattern used in deriving the reject option risk bound, but require us to first define conditional strength and conditional correlation. The unconditional strength and correlation used in the standard generalization bound and the reject option risk are based on the margin distribution, illustrated in Fig. 2. The conditional strengths and correlations required here are based on conditional score distributions illustrated in Figs. 3 and 4.

The correlations conditioned on the true label are defined as

$$\bar{\rho}_+ = \frac{2}{m(m-1)} \sum_{i \neq j} \mathrm{E}[\hat{y}_i(\mathbf{x})\hat{y}_j(\mathbf{x}) \mid y = +1] \tag{26}$$

$$\bar{\rho}_- = \frac{2}{m(m-1)} \sum_{i \neq j} \mathrm{E}[\hat{y}_i(\mathbf{x})\hat{y}_j(\mathbf{x}) \mid y = -1]. \tag{27}$$

The conditional strengths are defined as

$$s_+ = \mathrm{E}[\phi \mid y = +1] \tag{28}$$

$$s_- = -\mathrm{E}[\phi \mid y = -1]. \tag{29}$$

The conditional strength s_- is defined as the negative expected value of the conditional strength distribution so that s_- also equals $\mathrm{E}[z \mid y = -1]$. Thus,

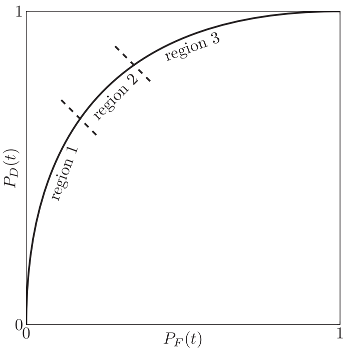


Fig. 5. Illustration of ROC split into Region 1: $t \in [s_+, 1]$, Region 2: $t \in [-s_-, s_+]$, and Region 3: $t \in [-1, -s_-]$.

$$s = s_+ \Pr[y = +1] + s_- \Pr[y = -1]. \tag{30}$$

The variance relationship of the margin distribution (17) also holds for the conditional score distributions:

$$\mathrm{var}(\phi \mid y = +1) \le \bar{\rho}_+\big(1 - s_+^2\big) \tag{31}$$

$$\mathrm{var}(\phi \mid y = -1) \le \bar{\rho}_-\big(1 - s_-^2\big). \tag{32}$$

We are now ready to bound the false alarm and detection probabilities, as well as the ROC.

We begin with the detection probability $P_D(t) = \Pr[\phi > t \mid y = +1]$. Using the Cantelli inequality (20) as before,

$$P_D(t) \le \frac{1}{1 + \frac{(\mathrm{E}[\phi|y=+1]-t)^2}{\mathrm{var}(\phi|y=+1)}}, \quad \mathrm{E}[\phi \mid y = +1] < t, \tag{33}$$

and using (28) and (31):

$$P_D(t) \le \frac{1}{1 + \frac{(s_+-t)^2}{\bar{\rho}_+(1-s_+^2)}}, \quad s_+ < t. \tag{34}$$

This is an upper bound for the tail $s_+ < t$. We may also get a lower bound for the tail $s_+ > t$ also using the Cantelli inequality, (28), and (31):

$$P_D(t) \ge \frac{1}{1 + \frac{\bar{\rho}_+(1-s_+^2)}{(s_+-t)^2}}, \quad s_+ > t. \tag{35}$$

The same calculations give us bounds on the tails of the false alarm probability $P_F(t) = \Pr[\phi > t \mid y = -1]$:

$$P_F(t) \le \frac{1}{1 + \frac{(s_-+t)^2}{\bar{\rho}_-(1-s_-^2)}}, \quad -s_- < t, \tag{36}$$

and

$$P_F(t) \ge \frac{1}{1 + \frac{\bar{\rho}_-(1-s_-^2)}{(s_-+t)^2}}, \quad -s_- > t. \tag{37}$$

The domains over which the bounds (34)-(37) are active delineate three intervals of the threshold:

1. $t \in [s_+, 1]$
2. $t \in [-s_-, s_+]$
3. $t \in [-1, -s_-]$,

which correspond to three regions of the ROC, as shown in Fig. 5. The bounds are given by region in Table 1. Region 1 corresponds to the low false alarm regime and Region 3 to the low missed detection regime. As we will show in Section 5, the bounds are predictive of empirical detection and false alarm rates.

TABLE 1
Generalization Bounds on False Alarm and Detection Probabilities

Region 1 $[s_+, 1]$	Region 2 $[-s_-, s_+]$	Region 3 $[-1, -s_-]$
$P_F \le \frac{1}{1+\frac{(s_-+t)^2}{\bar{\rho}_-(1-s_-^2)}}$	$P_F \le \frac{1}{1+\frac{(s_-+t)^2}{\bar{\rho}_-(1-s_-^2)}}$	$P_F \ge \frac{1}{1+\frac{\bar{\rho}_-(1-s_-^2)}{(s_-+t)^2}}$
$P_D \le \frac{1}{1+\frac{(s_+-t)^2}{\bar{\rho}_+(1-s_+^2)}}$	$P_D \ge \frac{1}{1+\frac{\bar{\rho}_+(1-s_+^2)}{(s_+-t)^2}}$	$P_D \ge \frac{1}{1+\frac{\bar{\rho}_+(1-s_+^2)}{(s_+-t)^2}}$

In Region 2, we have an upper bound for the ROC abscissa P_F, and a lower bound for the ROC ordinate P_D. Therefore, the two Region 2 bounds together constitute a lower bound for the ROC when the threshold is in Region 2. Examining the Region 2 bounds, it can be seen that at $t = -s_-$, the P_F upper bound (36) becomes 1, and that at $t = s_+$, P_D lower bound (35) becomes zero. If we extend the Region 2 P_F bound to be 1 for $t < -s_-$ and the Region 2 P_D bound to be zero for $t > s_+$, then we have an implicitly specified lower bound for the full ROC.

Let us express this lower bound for the full ROC explicitly as a function of P_F. Taking the Region 2 P_F bound (36) as an equality, we solve for the threshold t in terms of P_F:

$$t = \sqrt{\big(P_F^{-1} - 1\big)\big(\bar{\rho}_-\big(1 - s_-^2\big)\big)} - s_-. \tag{38}$$

Substituting this expression (38) for t into the Region 2 P_D bound (35) taken as an equality, we find the following lower bound for the ROC:

$$\mathrm{ROC} \ge \left[1 + \frac{\bar{\rho}_+\big(1 - s_+^2\big)}{\bar{\rho}_-\big(1 - s_-^2\big)}\left(\frac{s_+ + s_-}{\sqrt{\bar{\rho}_-\big(1 - s_-^2\big)}} - \sqrt{P_F^{-1} - 1}\right)^{-2}\right]^{-1} \tag{39}$$

for $t > s_+$ and zero otherwise.

Let us define the parameters

$$\eta_M = \frac{\bar{\rho}_+\big(1 - s_+^2\big)}{\big(s_+ + s_-\big)^2} \tag{40}$$

$$\eta_F = \frac{\bar{\rho}_-\big(1 - s_-^2\big)}{\big(s_+ + s_-\big)^2} \tag{41}$$

and simplify to obtain the ROC lower bound:

$$\mathrm{ROC} \ge \begin{cases} 0, & P_F \le \dfrac{\eta_F}{\eta_F + 1} \\ \frac{1}{1+\eta_M\left(1-\sqrt{\eta_F\left(P_F^{-1}-1\right)}\right)^{-2}}, & P_F > \dfrac{\eta_F}{\eta_F + 1} \end{cases}. \tag{42}$$

Note the functional similarity between η_M and η_F, and the zero-one generalization error bound (18).

For zero-one generalization error and reject option risk, we found that small $\bar{\rho}$ and large s are guidelines for good performance. Looking at the ROC, it is desired that for a given threshold, the false alarm probability be small and the detection probability be large. In the Region 2 bounds shown in Table 1, we see that for small false alarm probability, $\bar{\rho}_-$ should be small and s_- should be large. For large detection probability, $\bar{\rho}_+$ should be small and s_+ should be large. The guidelines in Region 2 for the conditional strengths and correlations are analogous to the guidelines for zero-one error and reject option risk, but may not be easily achieved in practice because large strength necessitates large correlation.

Notably, however, the guidelines in Regions 1 and 3 are not analogous. Examining the bounds in Table 1, we see that in Region 1, $\bar{\rho}_+$ should be large rather than small for large detection probability, and in Region 3, $\bar{\rho}_-$ should be large rather than small for small false alarm probability. As has been discussed earlier, Region 1 is the low false alarm regime and Region 3 is the low missed detection regime. Thus, improving detection probability in the low false alarm regime and improving false alarm probability in the low missed detection regime are not of most interest.

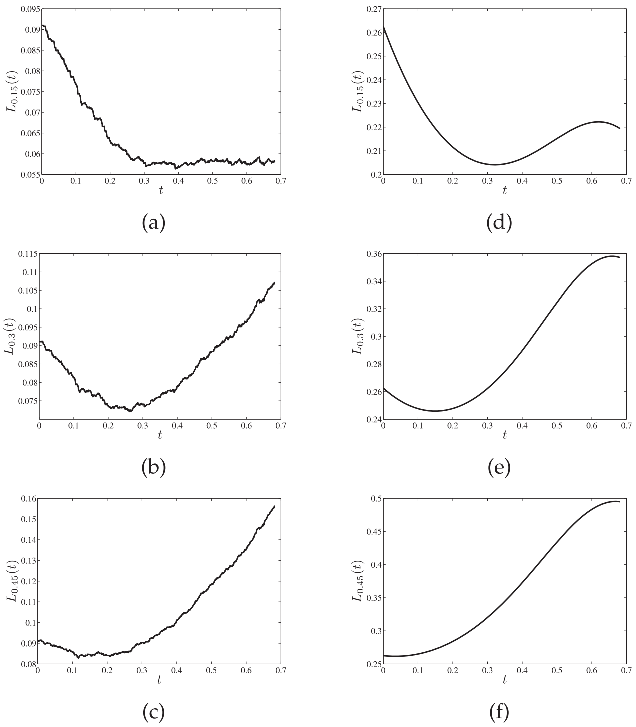
We see that the ROC lower bound function (42) goes to zero as P_F goes to $\eta_F/(\eta_F + 1)$ and is zero for small values as well. To push this function to the left, which corresponds to improvement in the ultralow false alarm regime, we would like η_F to be as close to zero as possible. In the definition of η_F (41), both conditional strengths appear, but only the negative conditional correlation appears. The guideline for ultralow false alarm probability from the definition of η_F is small $\bar{\rho}_-$ and large s_+ and s_-. Very large strength and very small correlation cannot be achieved simultaneously, but since $\bar{\rho}_+$ is not in the definition of η_F, we can make the correlation of the positive class high to get the strength of the positive class high for ultralow false alarms.

Correspondingly, to push the ROC up in the low missed detection regime, we would like η_M to be as close to zero as possible. In its definition (40), there is no negative class correlation. Thus, for ultralow missed detections, the guideline is to make the correlation of the negative class high to get the strength of the negative class high. Together, the guidelines for ultralow false alarms and ultralow missed detections are unlike the guidelines for zero-one loss. In these regimes, correlation is not to be kept small for one of the two classes: Base classifiers should not be fully diverse.

It is straightforward to derive bounds based on the Cantelli inequality for the combined formulation of different operating points and reject option presented in Section 3.4 in the same manner shown for the reject option and operating points formulations separately. What remains to be shown is that the bounds are predictive of empirical behavior. Providing guidelines on strength and correlation based on the bound functions is one thing. Showing that the guidelines transfer over to empirical behavior is another, and one we tackle in the following section.

5 Empirical Results

In this section, the similarity between the bounds derived in Section 4 and the empirical versions of those classification performance quantities are examined on real-world data sets from the UCI Machine Learning Repository [33]. Specifically, we look at the SPECTF heart data set in which the task is to classify patients as normal or abnormal using single proton emission computed tomography image features, and the Parkinsons data set in which the task is to diagnose Parkinson's disease from biomedical voice measurements. The SPECTF heart data set has 44 features and 267 samples. The Parkinsons data set has 22 features and 197 samples. The ensemble classifier that we consider is the random forest classifier [8]; we use the Matlab statistics toolbox implementation TreeBagger.



(a) (d)

(b) (e)

(c) (f)

Fig. 6. Comparison of (a)-(c) empirical and (d)-(f) analytical bound of reject option risk as a function of rejection threshold for three different values of rejection cost on the Parkinsons data set.

5.1 Reject Option Risk

We first examine the reject option risk using default parameter settings of TreeBagger.[1] Specifically, we look at the reject option risk as a function of the rejection threshold for different rejection cost values, and also at the risk-minimizing threshold as a function of cost.

We look at the Parkinsons data set first. In the medical diagnosis setting, it is useful for an automatic classification algorithm to have a reject option, allowing for further tests on difficult to classify patients. We train eleven random forests, each composed of 500 classification trees, with different random seeds on the data set, and obtain the out-of-bag margin distribution. We calculate the empirical reject option risk as a function of the rejection threshold and plot it in Fig. 6.[2]

For the different cost values shown, $c = 0.15, 0.30, 0.45$, the shape of the risk function is different. In particular, the

[1] Defaults include using the full input data to sample with replacement, using the square root of the number of features to select at random for each decision split, and having a minimum of one observation per tree leaf.

[2] All plots are mean values based on the 11 random forests with different random seeds.

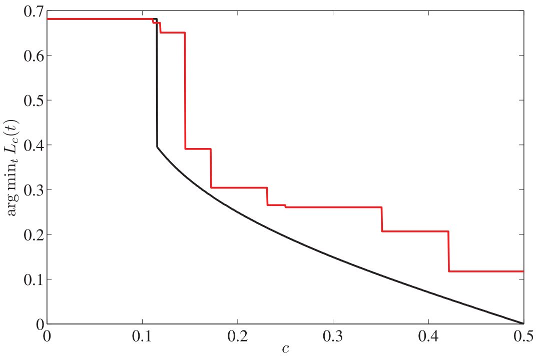


Fig. 7. Rejection threshold that minimizes risk as a function of rejection cost, empirically (red line) and on the analytical bound (black line) for the Parkinsons data set.

threshold that minimizes the risk is small, intermediate, and large for the respective costs. Fig. 6 also plots the risk bound derived in Section 4.2 for the different cost values. The risk bound functions mirror the empirical risk functions in shape. Additionally, the minimizing threshold of the bound is close to that of the empirical reject option risk.

To examine this further, we plot the minimizing threshold of the risk as a function of the cost in Fig. 7. This function, both the empirical and bound versions, is seen to be nonincreasing: The higher the cost of a rejection, the smaller the rejection region in feature space, cf. Fig. 1. The bound version jumps to s at a particular small value of c because the $L_c(t)$ function becomes monotonically decreasing in t at that value of c. It is especially enlightening that the minimizing threshold of the bound is quite predictive of the empirical minimizing threshold. Setting the rejection threshold is an important task in practice. Due to the predictive quality of the bound that has been derived in this paper, the bound may be used to set the threshold for a given cost value.

The second data set examined is also from the medical domain, specifically heart disease. For this data set also, 11 random forest classifiers containing 500 trees with different seeds are learned. We give the same plots for the SPECTF heart data set as for the Parkinsons data set in Figs. 8 and 9. The same features of the empirical risk and risk bound are seen, the most important of which is again that the bound may be used to set the rejection threshold.

Due to the small number of samples in both the Parkinsons and SPECTF data sets, there are few distinct risk-minimizing empirical thresholds in Figs. 7 and 9. As an example with a larger number of samples and distinct thresholds, we also give results for the spambase data set, in which the task is to determine whether an e-mail is spam: an unsolicited, commercial message. In the e-mail setting, it is useful for a spam filter to have a reject option, allowing the e-mail recipient the opportunity to decide whether a particular message that is difficult to classify is spam. The measurements upon which the spam filter makes its determination are 54 percentages that report the fraction of a message that matches a particular word or character, and three counts related to runs of capital letters. The data set contains 4,601 samples and we use 50 trees per random forest, again with 11 random forests.

Plots for the spambase data set corresponding to those for the Parkinsons and SPECTF data sets are given in Figs. 10 and 11. The match in structure between the empirical reject option risks and their bounds is striking here too, and the predictive quality of the bound on the risk-minimizing threshold is more apparent.

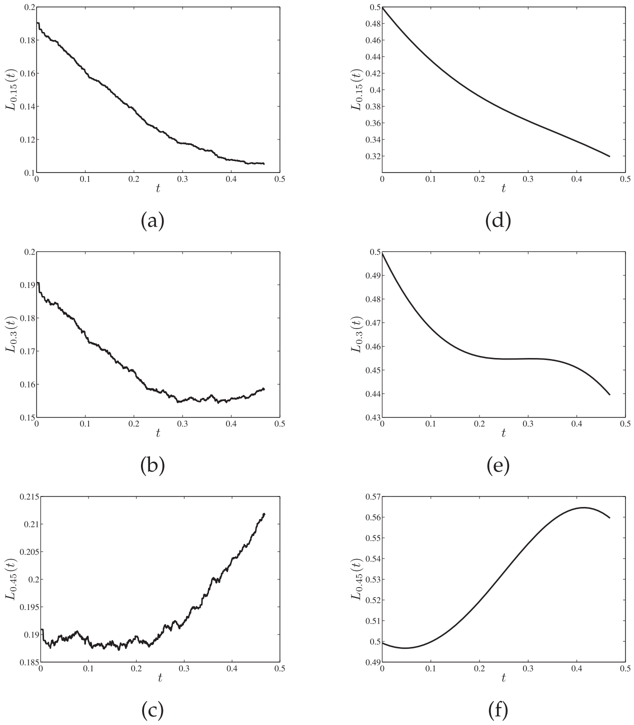


Fig. 8. Comparison of (a)-(c) empirical and (d)-(f) analytical bound of reject option risk as a function of rejection threshold for three different values of rejection cost on the SPECTF heart data set.

5.2 Receiver Operating Characteristic

Having seen the predictive quality of the derived bounds on reject option risk, we now turn to empirical validation of the ROC bounds. To investigate the ROC bounds empirically, we create different ensembles by varying the number of features that are randomly sampled per classification tree node in the random forest. As with the reject option, we use the default values from TreeBagger for all parameters of the random forest besides the number of features per node. Varying the number of sampled features changes the conditional strengths and correlations, as shown in Fig. 12 for the SPECTF heart data set. Consequently, the parameters

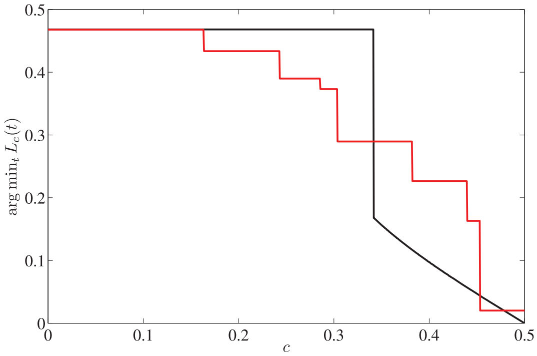
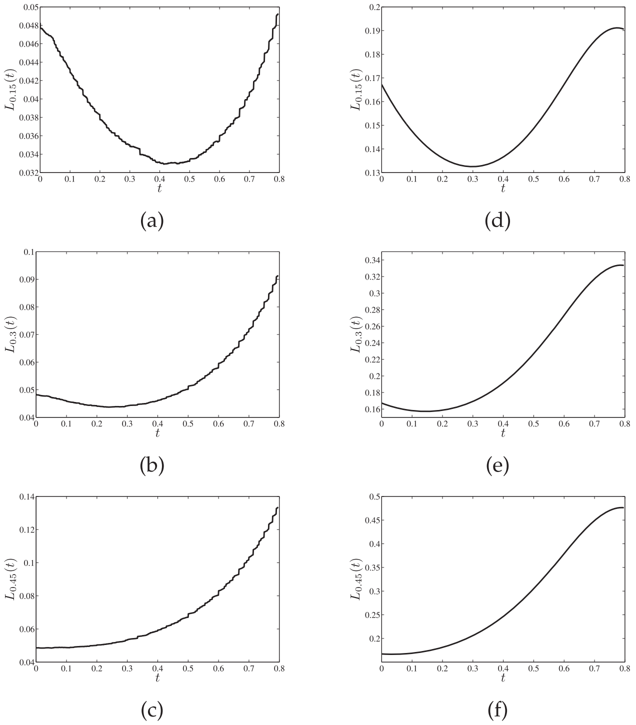


Fig. 9. Rejection threshold that minimizes risk as a function of rejection cost, empirically (red line) and on the analytical bound (black line) for the SPECTF heart data set.



(a) (d)

(b) (e)

(c) (f)

Fig. 10. Comparison of (a)-(c) empirical and (d)-(f) analytical bound of reject option risk as a function of rejection threshold for three different values of rejection cost on the spambase data set.

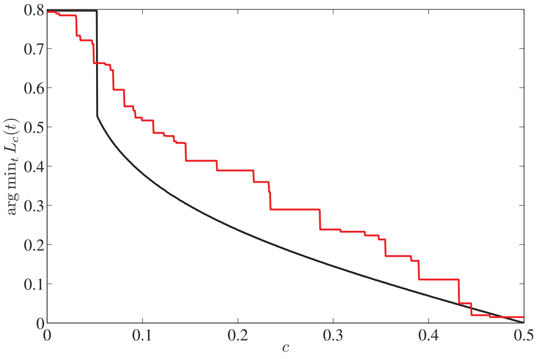
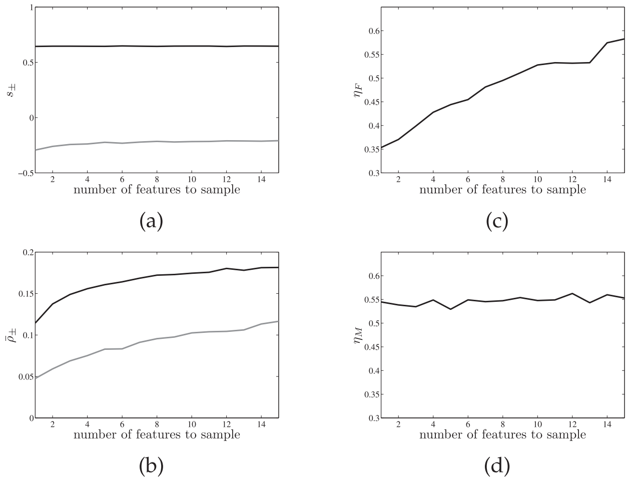


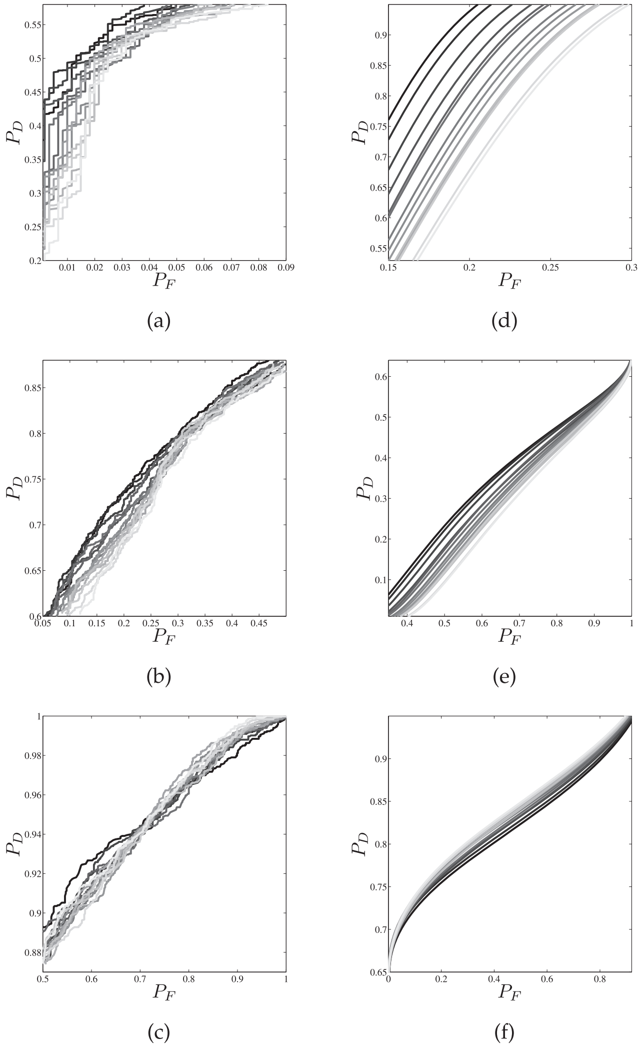
Fig. 11. Rejection threshold that minimizes risk as a function of rejection cost, empirically (red line) and on the analytical bound (black line) for the spambase data set.



(a) (c)

(b) (d)

Fig. 12. Conditional (a) strengths and (b) correlations (negative: gray line; positive: black line), (c) η_F, and (d) η_M as a function of the number of features per node in the random forest for the SPECTF heart data set.



(a) (d)

(b) (e)

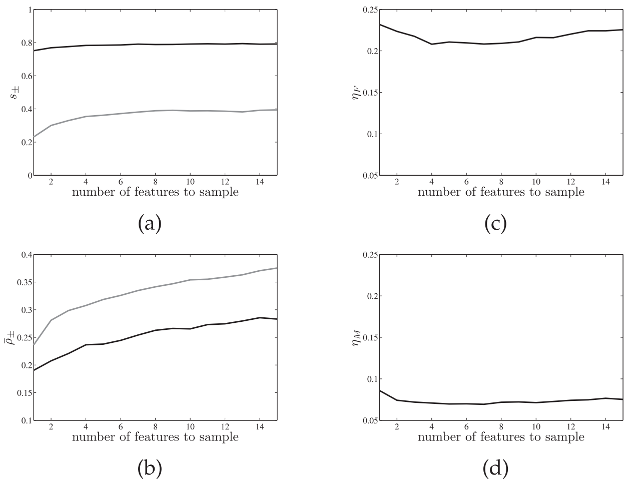
(c) (f)

Fig. 13. Comparison of (a)-(c) empirical ROC and (d)-(f) bound for different numbers of features per random forest node (black $= 1$, light gray $= 15$) in Regions 1 (a, d), 2 (b, e), and 3 (c, f) with the SPECTF heart data set.

η_F and η_M also change as a function of the number of sampled features.

In Fig. 13, we plot the empirical ROCs and ROC bounds for the three different regions delineated in Section 4.3. In Region 1, the low false alarm region, we see that the random forest that samples one feature per node provides the best performance and the performance is progressively worse as we increase the number of sampled features per node. This ordering in the empirical results is reproduced by the bound. In Region 2, it is again the random forest with one feature per node that is superior and 15 features per node that is inferior. The shape of the bound functions matches the shape of the empirical ROCs in this region quite well. In Region 3, the low missed detection region, we see that there is a crossover from the forest with one feature per node being best to it being worst, and the forest with 15 features per node going from the worst to best. This crossover occurs in the bound functions as well.

This crossover can be understood by examining η_F and η_M. For small numbers of sampled features, η_F is less than



(a) (c)

(b) (d)

Fig. 14. Conditional (a) strengths and (b) correlations (negative: gray line; positive: black line), (c) η_F, and (d) η_M as a function of the number of features per node in the random forest for the Parkinsons data set.

η_M, and the opposite for large numbers of sampled features. These parameters show that small numbers of sampled features are preferred in the low false alarm regime and large numbers of sampled features are preferred in the low missed detection regime.

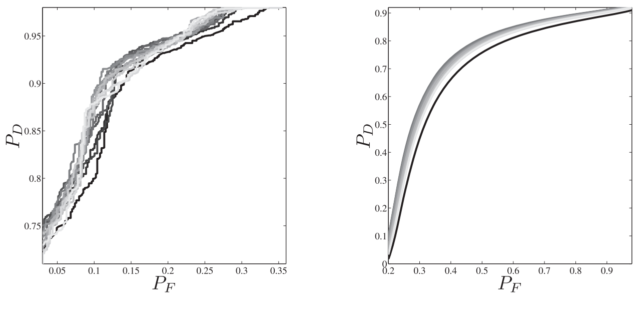
Fig. 14 shows the conditional strengths and correlations, η_F, and η_M as a function of the number of features sampled at each classification tree node for the Parkinsons data set. In these plots, it is seen that η_F and η_M are minimum at intermediate numbers of features sampled per node. The empirical ROC and ROC bound are shown in Fig. 15. For the Parkinsons data set, essentially the entire ROC is in Region 2 [11], so only the Region 2 ROC is shown.

In the figure, we can see that the ensemble with one feature per node is inferior, but also that the ensemble with 15 features per node is not the best. An intermediate random forest provides the best performance. This ordering is reflected in the bound functions as well and is also reflective of η_F and η_M being minimum at intermediate values.

In the empirical examples given here, we see qualitatively that the bounds are quite predictive of the relative performance of ensembles having different conditional strengths and correlations. This predictive quality is quantified by Prenger et al. [11]; a similarity metric is defined by computing the correlation coefficient between the empirical P_F and its bound across the varying number of features sampled per node of the classification tree in the random forest at 10,000 uniformly distributed threshold values between -1 and $+1$. Similarly, the correlation coefficient is calculated between the empirical P_D and its bound. The average of these 20,000 correlation coefficients is taken as the similarity and is repeated for 101 random forests with different random seeds. The similarity is found to be 0.6676 for the SPECTF heart data set and 0.5004 for the Parkinsons data set, both with statistical significance of level 0.01.

6 Conclusion

In this paper, we derive generalization bounds on the reject option risk, probability of false alarm, probability of missed detection, and ROC for ensemble classifiers. These bounds are derived using the Cantelli inequality based on interpretable statistics of the ensemble margin distribution, the strength and correlation. The bounds are not tight in absolute value, but nevertheless have predictive value.



(a) (b)

Fig. 15. Comparison of (a) empirical ROC and (b) bound for different numbers of features per random forest node (black $= 1$, light gray $= 15$) in Region 2 with the Parkinsons data set.

As we show on real-world data sets, bound functions mirror their empirical counterparts in structure and thus are useful to set rejection thresholds, rank ensembles with different characters, and provide guidelines for ensemble choice. Somewhat counterintuitive guidelines are revealed in the ultralow false alarm and ultralow missed detection regimes: Correlation conditioned on the positive class should be increased without penalty to improve false alarm probability and correlation conditioned on the negative class should be increased without penalty to improve detection probability. This is in contrast to the typical guideline of desiring small correlation, which is what is desired for zero-one loss, reject option risk, and classification performance in Region 2.

In future work, it would be of interest to extend the ROC analysis to allow us to use the derived bounds in setting the classification threshold in a Neyman-Pearson setting [7]. It would also be of interest to develop generalization error bounds for ensemble classifiers used to learn severely imbalanced data [34] and connect such bounds to the analysis of [29].

Acknowledgments

Part of this work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, no. 2, pp. 123-140, Aug. 1996.

[2] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Computer and System Sciences,* vol. 55, no. 1, pp. 119-139, Aug. 1997.

[3] L. Breiman, "Arcing Classifiers," *Annals of Statistics,* vol. 26, no. 3, pp. 801-849, June 1998.

[4] Y. Zhang and W.N. Street, "Bagging with Adaptive Costs," *IEEE Trans. Knowledge Data Eng.,* vol. 20, no. 5, pp. 577-588, May 2008.

[5] K.R. Varshney, D. Fang, A.R. Heching, A. Mojsilović, and M. Singh, "Classification of IT Service Tickets for Defect Prevention," *Proc. INFORMS Ann. Meeting,* Nov. 2011.

[6] C.K. Chow, "On Optimum Recognition Error and Reject Tradeoff," *IEEE Trans. Information Theory,* vol. IT-16, no. 1, pp. 41-46, Jan. 1970.

[7] H.L. Van Trees, *Detection, Estimation, and Modulation Theory.* Wiley, 1968.

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

[9] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *Annals of Statistics,* vol. 26, no. 5, pp. 1651-1686, Oct. 1998.

[10] L. Wang, M. Sugiyama, Z. Jing, C. Yang, Z.-H. Zhou, and J. Feng, "A Refined Margin Analysis for Boosting Algorithms via Equilibrium Margin," *J. Machine Learning Research,* vol. 12, pp. 1835-1863, June 2011.

[11] R.J. Prenger, T.D. Lemmond, K.R. Varshney, B.Y. Chen, and W.G. Hanley, "Class-Specific Error Bounds for Ensemble Classifiers," *Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining,* pp. 843-852, July 2010.

[12] K.R. Varshney, "A Risk Bound for Ensemble Classification with a Reject Option," *Proc. IEEE Statistical Signal Processing Workshop,* pp. 773-776, June 2011.

[13] H. Kargupta, B.-H. Park, and H. Dutta, "Orthogonal Decision Trees," *IEEE Trans. Knowledge Data Eng.,* vol. 18, no. 8, pp. 1028-1042, Aug. 2006.

[14] L.L. Minku, A.P. White, and X. Yao, "The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift," *IEEE Trans. Knowledge Data Eng.,* vol. 22, no. 5, pp. 730-742, May 2010.

[15] L.L. Minku and X. Yao, "DDD: A New Ensemble Approach for Dealing with Concept Drift," *IEEE Trans. Knowledge Data Eng.,* vol. 24, no. 4, pp. 619-633, Apr. 2012.

[16] B. Verma and A. Rahman, "Cluster-Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning," *IEEE Trans. Knowledge Data Eng.,* vol. 24, no. 4, pp. 605-618, Apr. 2012.

[17] N. Usunier, M.-R. Amini, and P. Gallinari, "A Data-Dependent Generalisation Error Bound for the AUC," *Proc. ICML Workshop ROC Analysis in Machine Learning,* Aug. 2005.

[18] R. Herbei and M.H. Wegkamp, "Classification with Reject Option," *Canadian J. Statistics,* vol. 34, no. 4, pp. 709-721, Dec. 2006.

[19] P.L. Bartlett and M.H. Wegkamp, "Classification with a Reject Option Using a Hinge Loss," *J. Machine Learning Research,* vol. 9, pp. 1823-1840, Aug. 2008.

[20] C. Scott, G. Bellala, and R. Willett, "The False Discovery Rate for Statistical Pattern Recognition," *Electronic J. Statistics,* vol. 3, pp. 651-677, 2009.

[21] M. Yuan and M.H. Wegkamp, "Classification Methods with Reject Option Based on Convex Risk Minimization," *J. Machine Learning Research,* vol. 11, pp. 111-130, Jan. 2010.

[22] R. El-Yaniv and Y. Wiener, "On the Foundations of Noise-Free Selective Classification," *J. Machine Learning Research,* vol. 11, pp. 1605-1641, May 2010.

[23] O. Bousquet, "New Approaches to Statistical Learning Theory," *Annals Inst. of Statistical Math.,* vol. 55, no. 2, pp. 371-389, June 2003.

[24] Y. Freund, Y. Mansour, and R.E. Schapire, "Generalization Bounds for Averaged Classifiers," *Annals of Statistics,* vol. 32, no. 4, pp. 1698-1722, Aug. 2004.

[25] G. Biau, L. Devroye, and G. Lugosi, "Consistency of Random Forests and Other Averaging Classifiers," *J. Machine Learning Research,* vol. 9, pp. 2015-2033, Nov. 2008.

[26] L.I. Kuncheva and C.J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning,* vol. 51, no. 2, pp. 181-207, May 2003.

[27] G.I. Webb and Z. Zheng, "Multistrategy Ensemble Learning: Reducing Error by Combining Ensemble Learning Techniques," *IEEE Trans. Knowledge Data Eng.,* vol. 16, no. 8, pp. 980-991, Aug. 2004.

[28] E.K. Tang, P.N. Suganthan, and X. Yao, "An Analysis of Diversity Measures," *Machine Learning,* vol. 65, no. 1, pp. 247-271, Oct. 2006.

[29] S. Wang and X. Yao, "Relationships between Diversity of Classification Ensembles and Single-Class Performance Measures," *IEEE Trans. Knowledge Data Eng.,* vol. 25, no. 1, pp. 206-219, Jan. 2013.

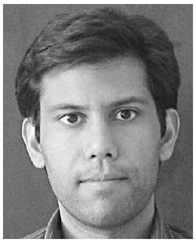
[30] T. Pietraszek, "On the Use of ROC Analysis for the Optimization of Abstaining Classifiers," *Machine Learning,* vol. 68, no. 2, pp. 137-169, Aug. 2007.

[31] F.P. Cantelli, "Sui Confini Della Probabilità," *Atti del Congresso Internazionale dei Matematici,* vol. 6, pp. 47-59, Sept. 1928.

[32] D. Bertsimas and I. Popescu, "Optimal Inequalities in Probability Theory: A Convex Optimization Approach," *SIAM J. Optimization,* vol. 15, no. 3, pp. 780-804, 2005.

[33] A. Asuncion and D.J. Newman, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml, 2007.

[34] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," Technical Report 666, Dept. of Statistics, Univ. of California, Berkeley, July 2004.



Kush R. Varshney received the BS degree (magna cum laude) in electrical and computer engineering with honors from Cornell University, Ithaca, New York, in 2004. He received the SM and PhD degrees, both in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2006 and 2010, respectively. He is a research staff member in the Business Analytics and Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York. While at MIT, he was a research assistant with the Stochastic Systems Group in the Laboratory for Information and Decision Systems and a National Science Foundation Graduate Research fellow. He has been a visiting student with École Centrale, Paris, and an intern with Lawrence Livermore National Laboratory, Sun Microsystems, and Sensis Corporation. His research interests include statistical signal processing, statistical learning, and image processing. He is a member of Eta Kappa Nu and Tau Beta Pi. He received a Best Student Paper Travel Award at the 2009 International Conference on Information Fusion. He received an IBM Research Division Award for business impact of outsourcing analytics and an IBM Outstanding Technical Achievement Award for contributions to salesforce analytics, both in 2012. He is a member of the IEEE.

Ryan J. Prenger received the PhD degree in physics from the University of California, Berkeley, where he developed machine learning algorithms for the encoding and decoding of neural responses to natural images. He is currently a research engineer at the Lawrence Livermore National Laboratory developing and applying machine learning approaches to classification and anomaly detection problems. He has been an author on several research papers in various journals and conferences including KDD, *Nature*, and *Neuron*.

Tracy L. Marlatt received the master's degree in mathematical sciences from Clemson University. In 2002, she joined Lawrence Livermore National Laboratory, where she focused on the development and application of new machine learning technologies in classification and anomaly detection. She has led several research projects in these and related tasks, has published more than 20 research papers in journals and conferences, and was granted a patent resulting directly from her work.

Barry Y. Chen received the PhD degree in electrical engineering and computer sciences from the University of California, Berkeley, where he developed new neural network learning algorithms for automatic speech recognition. He is currently a research engineer at the Lawrence Livermore National Laboratory developing and applying machine learning approaches to classification and anomaly detection problems. He has published more than 20 research papers in various journals and conferences.

William G. Hanley received the BS degree in mathematics with honors from the University of California, Riverside, in 1978. He received the MS degree in computer science from the University of California, Los Angeles, in 1983. He received the MS and PhD degrees, both in applied statistics from the University of California, Riverside, in 1993 and 1998, respectively. He is a senior managing scientist within the Technology Development Practice of Exponent Inc., Menlo Park, California. Previously, while at Lawrence Livermore National Laboratory, he served as the deputy division leader for the National Security Engineering Division, section leader for the Systems and Intelligence Analysis Section, and the focus area leader for Engineering Systems for Knowledge and Inference. His research interests include statistical learning, decision methodologies, and system modeling. He is a member of the American Statistical Association.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.