A RISK BOUND FOR ENSEMBLE CLASSIFICATION WITH A REJECT OPTION

Kush R. Varshney

Business Analytics and Mathematical Sciences Department, IBM Thomas J. Watson Research Center
1101 Kitchawan Rd., Route 134, Yorktown Heights, NY 10598, USA

ABSTRACT

Signal classification is an important task in numerous application domains that is increasingly being approached through ensemble methods, such as those involving boosting and bootstrap aggregation. In decision support scenarios, it is often of interest for automatic classification algorithms to abstain from making decisions on the most uncertain signals; this is known as classification with a reject option. In this work, a bound on generalization error for ensemble classification with a reject option is derived that involves two intuitive properties of the ensemble: average strength and mean correlation. The bound is shown to be predictive of empirical classification behavior and useful in setting the rejection threshold for a given rejection cost.

Index Terms— ensemble classifier, random forest, reject option, generalization bound

1. INTRODUCTION

Signal classification is a common task in applications ranging from surveillance to medicine to communication. Automatic classification algorithms may be used to support human decision makers and reduce their load. For example, a classification algorithm may be used to make a diagnosis based on a blood test with little or no input from a physician. In such decision support scenarios, it is of interest that the algorithm abstain from making a classification on the most uncertain signals rather than making an incorrect classification. Such abstention is known as *classification with a reject option* [1, 2]. Depending on the application, the unclassified, or rejected, signals may be ignored or may be further examined by various means, including manual inspection or obtaining additional measurements or features to reduce uncertainty, akin to sequential hypothesis testing [3].

Over the last decade, ensemble classifiers, including random forests [4], have proven to be powerful, robust and scalable, and have become first choice classification algorithms among many signal analysts and data mining practitioners. The main idea of ensemble classification is to train many diverse, but perhaps weak, base classifiers and have them vote to determine an overall classification.

Bounds on the generalization error — the probability of classification error on signals apart from those used to train the decision rule — are often useful in understanding the properties and behaviors of classification algorithms. A highly interpretable generalization bound for ensemble classification involving two properties of the collection of base classifiers, *average strength* and *mean correlation*, is developed in [4]. This generalization bound's primary utility is in the intuition it provides regarding the influence of ensemble properties on performance, but it also has a strong correlation with empirical classification error on new signals not used in training.

In this work, a risk bound for ensemble classification with a reject option is developed that also depends on the quantities average strength and mean correlation. Here also, the bound provides intuition about the factors that go into the classification performance of the ensemble. The main free parameter in classification with a reject option is the rejection threshold. In addition to the intuition provided by the risk bound that is developed, it is seen that the value of the threshold that minimizes the bound is approximately equal to the value that minimizes the empirical risk on real-world data sets. Thus the bound may be used to guide the selection of the threshold. The derivation of the reject option risk bound in this paper closely follows the derivation of generalization bounds for receiver operating characteristics of ensemble classifiers presented in [5].

An overview of ensemble classification at the depth required for the remainder of the paper is provided in Section 2. Implementation of the reject option and the definition of risk to measure classification performance are also discussed in that section. Making use of the Cantelli inequality, the risk is bounded in Section 3. The efficacy of the bound is shown in Section 4 through empirical comparison. Section 5 concludes.

2. ENSEMBLE CLASSIFICATION WITH A REJECT OPTION

Consider a classification problem in which class labels $y \in \{-1, +1\}$ are to be predicted using signals $\mathbf{x} \in \mathcal{X}$. The decision rule $\hat{y}(\cdot) : \mathcal{X} \to \{-1, +1\}$ is learned from labeled i.i.d. training data and applied to new, unseen and unlabeled test signals from the same distribution. In the specific case of ensemble classification, $\hat{y}$ is composed of base classifiers $\hat{y}_i(\cdot) : \mathcal{X} \to \{-1, +1\}$, $i = 1, \dots, m$. 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 positive then $\hat{y} = +1$, i.e. $\hat{y}(\mathbf{x}) = \operatorname{sign}(\phi(\mathbf{x}))$. In the context of classification with a reject option, the overall decision is the following:

$$\hat{y}(\mathbf{x}) = \begin{cases} -1, & \phi(\mathbf{x}) \leq -t \\ \text{reject}, & -t < \phi(\mathbf{x}) < t \\ +1, & \phi(\mathbf{x}) \geq t \end{cases} \tag{2}$$

where $t \geq 0$ is a rejection threshold. Decisions are most uncertain near the boundary between the two classes, which occurs at $\phi(\mathbf{x}) = 0$. Rejections are declared in the most uncertain regions of the signal space $\mathcal{X}$, where $\phi(\mathbf{x})$ is close to zero. 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.

978-1-4577-0568-7/11/$26.00 ©2011 IEEE

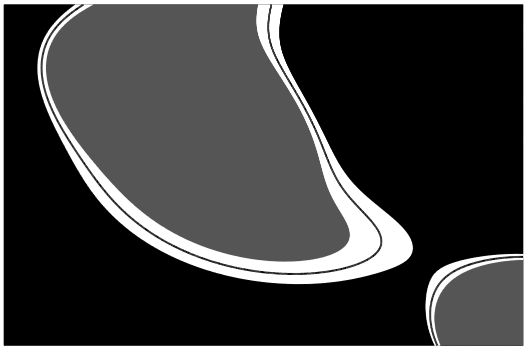


Fig. 1. Illustration of decision regions in signal space. The region $\hat{y} = +1$ is black, the region $\hat{y} = -1$ is gray, the region $\hat{y} =$ reject is white, and the boundary $\phi = 0$ is the black line.

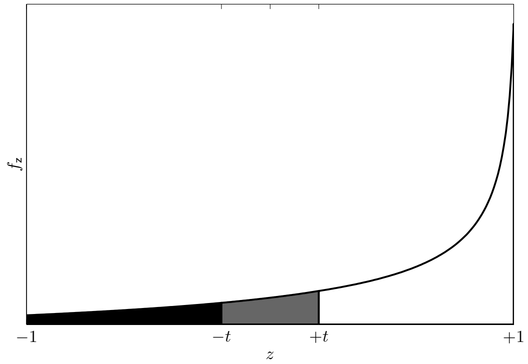


Fig. 2. Illustration of margin distribution $f_{\mathsf{z}}(z)$ marked with rejection threshold t. The area of the black region is the error probability P_E. The area of the gray region is the rejection probability P_R.

Define the margin to be $\mathsf{z} = \mathsf{y}\phi \in [-1, +1]$. Due to the special encoding $\mathsf{y} \in \{-1, +1\}$, the margin is negative for incorrect classifications and positive for correct classifications. With a reject option, 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 $f_{\mathsf{z}}(z)$. Thus, the probability of error

$$\begin{aligned} P_E &= \Pr[\mathsf{z} \leq -t] \\ &= \int_{-1}^{-t} f_{\mathsf{z}}(z) dz \end{aligned} \tag{3}$$

and the probability of rejection

$$\begin{aligned} P_R &= \Pr[-t < \mathsf{z} < t] \\ &= \int_{-t}^{t} f_{\mathsf{z}}(z) dz. \end{aligned} \tag{4}$$

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

As discussed in [2], a useful measure of performance is the reject option risk:

$$\begin{aligned} L_{c,t} &= P_E + cP_R \\ &= \Pr[\mathsf{z} \leq -t] + c\Pr[-t < \mathsf{z} < t] \end{aligned} \tag{5}$$

where the cost of error is one, the cost of correct classification is zero, and the cost of rejection is $0 \leq c \leq 1/2$. The reject option risk $L_{c,t}$ should be small for good performance. This risk is bounded in the following section.

3. BOUND FOR REJECT OPTION RISK

A bound for the risk $L_{c,t}$ involving the average strength and mean correlation of the ensemble is derived in this section. The average strength s measures the quality of the individual base classifiers and is the expected value of the margin, i.e. $s = \mathrm{E}[\mathsf{z}]$. The mean correlation $\bar{\rho}$ measures the diversity of the base classifiers and is the average pairwise correlation between base classifier outputs; it is shown in [4] that

$$\mathrm{var}(\mathsf{z}) \leq \bar{\rho}(1 - s^2). \tag{6}$$

It may be observed in Fig. 2 that

$$L_{c,t} = (1 - c)P_E + c\Pr[\mathsf{z} < t]. \tag{7}$$

The probability of error term is bounded first, followed by the $\Pr[\mathsf{z} < t]$ term; the two are then combined. The term $P_E = \Pr[\mathsf{z} \leq -t]$ is to be bounded using average strength and mean correlation. The Cantelli (one-sided Chebyshev) inequality is used toward this end [6]:

$$\Pr[\mathsf{z} - \mathrm{E}[\mathsf{z}] \leq -k] \leq \frac{1}{1 + \frac{k^2}{\mathrm{var}(\mathsf{z})}}, \quad k > 0. \tag{8}$$

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

$$\Pr[\mathsf{z} \leq -t] \leq \frac{1}{1 + \frac{(\mathrm{E}[\mathsf{z}]+t)^2}{\mathrm{var}(\mathsf{z})}}, \quad \mathrm{E}[\mathsf{z}] > -t, \tag{9}$$

and

$$P_E \leq \frac{1}{1 + \frac{(s+t)^2}{\bar{\rho}(1-s^2)}}, \quad s > -t, \tag{10}$$

due to (6) and the definition of s. It is a standard requirement in ensemble classification that the base classifiers have accuracy greater than random guessing, which translates to $s > 0$. Therefore $s > -t$ is assumed true and this constraint need not be further considered.

Now turning to the second half of the reject option risk expression (7), again by the Cantelli inequality,

$$\Pr[\mathsf{z} < t] \leq \frac{1}{1 + \frac{(\mathrm{E}[\mathsf{z}]-t)^2}{\mathrm{var}(\mathsf{z})}}, \quad \mathrm{E}[\mathsf{z}] > t \tag{11}$$

and also

$$\Pr[\mathsf{z} < t] \leq \frac{1}{1 + \frac{(s-t)^2}{\bar{\rho}(1-s^2)}}, \quad s > t. \tag{12}$$

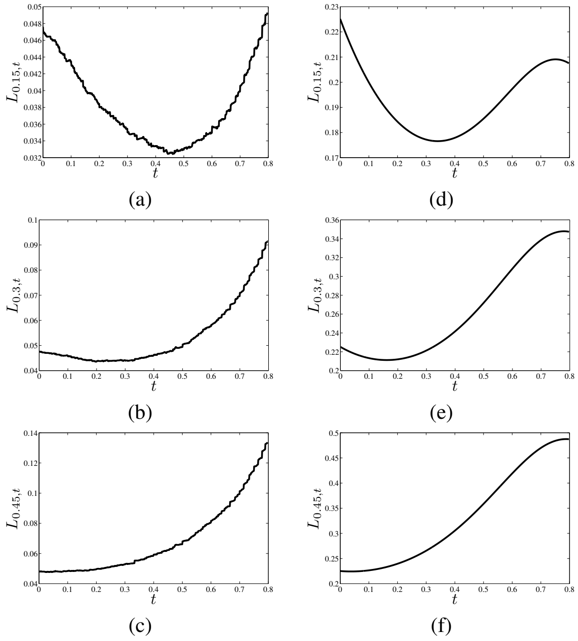
A bound for the reject option risk, the main result, is found by combining (10) and (12):

$$L_{c,t}(\mathsf{z}) \leq \frac{1-c}{1 + \frac{(s+t)^2}{\bar{\rho}(1-s^2)}} + \frac{c}{1 + \frac{(s-t)^2}{\bar{\rho}(1-s^2)}}, \quad s > t. \tag{13}$$

This bound is applicable when the rejection threshold is set below the average 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 4 show that this bound, although not tight in difference, is 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 (13) 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. 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 average strength and mean 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 $L_{c,t}(z)$ bound with respect to s is always negative, so the guideline is to have strength as high as possible. The derivative of the $L_{c,t}(z)$ 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. In practice, there are various ways to manipulate the strength and correlation of an ensemble [5], but it should be noted that the two are at odds with each other.



(a)

(d)

(b)

(e)

(c)

(f)

Fig. 3. 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.

4. EMPIRICAL EVIDENCE

In this section, the similarity between the reject option risk bound derived in the previous section and the empirical reject option risk is examined on real-world data sets from the UCI Machine Learning Repository [7]. The ensemble classifier that is considered is the random forest classifier [4]; the Matlab statistics toolbox implementation TreeBagger with default parameter settings is used.

The first data set examined is the spambase data set, in which the task is to determine whether an email is an unsolicited, commercial message, i.e. spam. Spam emails include product advertisements, pornography, chain letters, and messages purporting get-rich-quick schemes, whereas non-spam emails include personal and work-related messages. In the email setting, it is useful for a spam filter to have a reject option, allowing the email 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 4601 samples.

Eleven random forests with different random seeds, each composed of fifty classification trees, are learned from the data set, and the out-of-bag margin distribution is obtained. The empirical reject option risk is calculated as a function of the rejection threshold and is plotted in Fig. 3.[1] For the different cost values considered, $c = 0.15$, 0.30, 0.45, the shape of the risk function is different. In particular, the threshold that minimizes the risk is small, intermediate, and large for the respective costs. Fig. 3 also plots the risk bound derived in Section 3 for the different cost values. The risk bound functions mirror the empirical risk functions in shape. Additionally, the minimizing threshold of the bound nearly matches that of the empirical reject option risk.

To examine this further, the minimizing threshold of the risk is plotted as a function of the cost in Fig. 4. 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 signal 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.

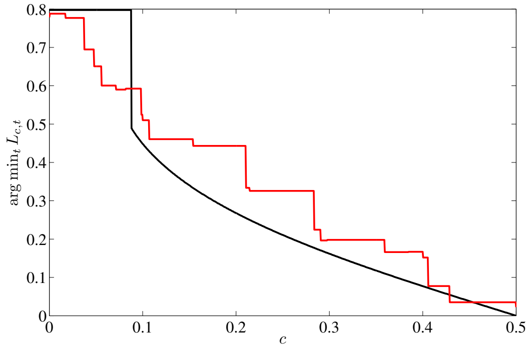


Fig. 4. Rejection threshold that minimizes risk as a function of rejection cost, empirically (red line) and on the analytical bound (black line) for the spambase data set.

[1] All plots are median values based on the eleven random forests with different random seeds.

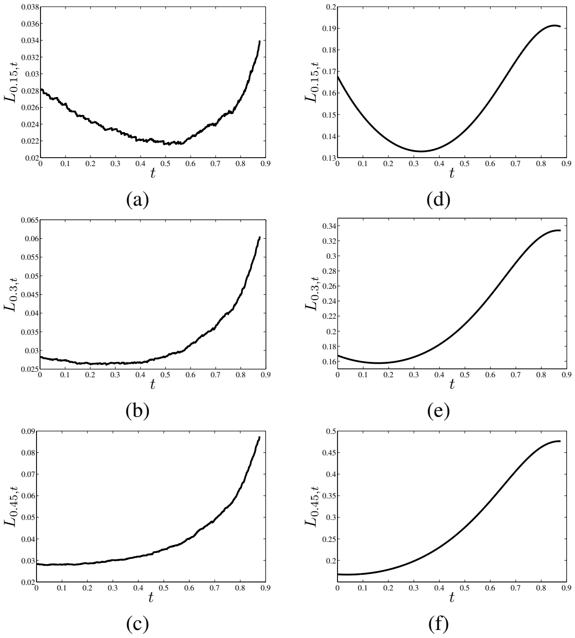


Fig. 5. Comparison of (a)–(c) empirical and (d)–(f) analytical bound of reject option risk as a function of rejection threshold for three different values of rejection cost on the internet advertisements data set.

The second data set examined is the internet advertisements data set, also from the UCI repository. Here, the task is to determine whether an image on a web page is an advertisement or not, based on 1558 different dimensions of measurements. These dimensions include the geometry of the image, and phrases in the uniform resource locator of the page and image, alternative text, anchor text and surrounding words. The advertisements data set contains 3279 samples.

For this data set also, eleven random forest classifiers with different seeds are learned, here with thirty-five trees per forest. The same plots as for the spambase data set are given for the internet advertisements data set in Fig. 5 and Fig. 6. 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.

5. CONCLUSION

Ensemble classification with a reject option has been analyzed in this paper. A bound for the reject option risk involving the average strength and mean correlation of the ensemble has been derived using the Cantelli inequality. The bound reveals the guidelines that, in analogy form, diverse (low correlation) and informed (high strength) committees of voters are good. Risk bounds in statistical learning theory are nearly always only useful in providing such guidelines; they are not objects to be meaningfully optimized [8]. However, the bound developed herein is unlike typical statistical learning theory bounds: the rejection threshold that minimizes the bound is a meaningful approximation to the optimal empirical rejection threshold. Thus, the contribution of the paper is beyond just theoretical interest, there is practical significance as well.

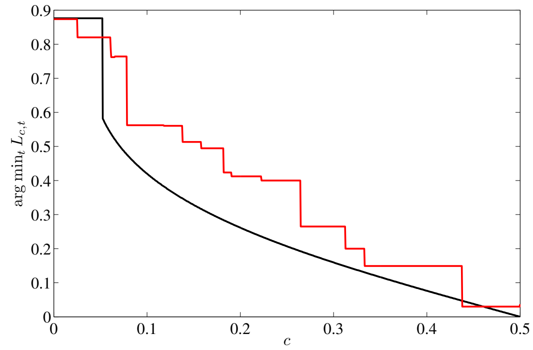


Fig. 6. Rejection threshold that minimizes risk as a function of rejection cost, empirically (red line) and on the analytical bound (black line) for the internet advertisements data set.

The bound developed here makes use of first and second order moments of the margin distribution along with the Cantelli inequality; future work may consider higher-order moments and inequalities such as those described in [9], although higher-order moments do not have the nice interpretability of average strength and mean correlation. Also, the analysis here is limited to binary classification, primarily due to notational convenience; it is straightforward to extend the analysis to multicategory classification using the definition of margin for the multicategory case given in [4]. In multicategory classification, an interesting extension of this work would be to consider the opposite of the reject option, akin to list decoding [10], in which more than one label may be assigned to a signal.

6. REFERENCES

[1] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 1, pp. 41–46, Jan. 1970.

[2] P. L. Bartlett and M. H. Wegkamp, "Classification with a reject option using a hinge loss," *J. Mach. Learn. Res.*, vol. 9, pp. 1823–1840, Aug. 2008.

[3] A. Wald, *Sequential Analysis*. New York: Wiley, 1947.

[4] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[5] R. J. Prenger, T. D. Lemmond, K. R. Varshney, B. Y. Chen, and W. G. Hanley, "Class-specific error bounds for ensemble classifiers," in *Proc. ACM SIGKDD Conf. Knowl. Disc. Data Min.*, Arlington, VA, Jul. 2010, pp. 843–852.

[6] F. P. Cantelli, "Sui confini della probabilità," in *Atti del Congresso Internazionale dei Matematici*, vol. 6, Bologna, Italy, Sep. 1928, pp. 47–59.

[7] A. Asuncion and D. J. Newman, "UCI machine learning repository," Available: http://archive.ics.uci.edu/ml, 2007.

[8] O. Bousquet, "New approaches to statistical learning theory," *Ann. Inst. Statist. Math.*, vol. 55, no. 2, pp. 371–389, Jun. 2003.

[9] D. Bertsimas and I. Popescu, "Optimal inequalities in probability theory: A convex optimization approach," *SIAM J. Optim.*, vol. 15, no. 3, pp. 780–804, 2005.

[10] G. D. Forney, Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 2, pp. 206–220, Mar. 1968.