Journal of Machine Learning Research 9 (2008) 171-174 Published 2/08

Response to Mease and Wyner, Evidence Contrary to the Statistical View of Boosting, *JMLR 9*:131–156, 2008

Yoav Freund YFREUND@UCSD.EDU
Department of Computer Science and Engineering
University of California
San Diego, CA 92093

Robert E. Schapire SCHAPIRE@PRINCETON.EDU
Princeton University
Department of Computer Science
35 Olden Street
Princeton, NJ 08540

Editor: Yoav Freund

For such a simple algorithm, it is fascinating and remarkable what a rich diversity of interpretations, views, perspectives and explanations have emerged of AdaBoost. Originally, AdaBoost was proposed as a "boosting" algorithm in the technical sense of the word: given access to "weak" classifiers, just slightly better in performance than random guessing, and given sufficient data, a true boosting algorithm can provably produce a combined classifier with nearly perfect accuracy (Freund and Schapire, 1997). AdaBoost has this property, but it also has been shown to be deeply connected with a surprising range of other topics, such as game theory, on-line learning, linear programming, logistic regression and maximum entropy (Breiman, 1999; Collins et al., 2002; Demiriz et al., 2002; Freund and Schapire, 1996, 1997; Kivinen and Warmuth, 1999; Lebanon and Lafferty, 2002). As we discuss further below, AdaBoost can been seen as a method for maximizing the "margins" or confidences of the predictions made by its generated classifier (Schapire et al., 1998). The current paper by Mease and Wyner, of course, focuses on another perspective, the so-called statistical view of boosting. This interpretation, particularly as expounded by Friedman et al. (2000), focuses on the algorithm as a stagewise procedure for minimizing the exponential loss function, which is related to the loss minimized in logistic regression, and whose minimization can be viewed, in a certain sense, as providing estimates of the conditional probability of the label.

Taken together, these myriad interpretations of AdaBoost form a robust theory of the algorithm that provides understanding from an extraordinary range of points of view in which each perspective tells us something unique about the algorithm. The statistical view, for instance, has been of tremendous value, allowing for the practical conversion of AdaBoost's predictions into conditional probabilities, as well as the algorithm's generalization and extension to many other loss functions and learning problems.

Still, each perspective has its weaknesses, which are important to identify to keep our theory in touch with reality. The current paper is superb in exposing empirical phenomena that are apparently difficult to understand according to the statistical view. From a theoretical perspective, the statistical interpretation has other weaknesses. As discussed by Mease and Wyner, this interpretation does not explain AdaBoost's observed tendency not to overfit, particularly in the absence of regularization

or early stopping. It also says little about how well AdaBoost will generalize when provided with a finite data set, nor how its ability to generalize is dependent on the complexity or simplicity of the base classifiers, an issue that arises in the experiments comparing decision stumps and decision trees in this role.

Much of the difficulty arises from the fact that AdaBoost is a *classification* algorithm (at least as it is used and studied in the current paper). This means that AdaBoost's purpose is to find a rule h that, given X, predicts one of the labels $h(X)$, and that attempts to achieve minimal probability of an incorrect classification (in which $h(X)$ disagrees with the true label Y). This is quite different from the problem of estimating the conditional probability $P(Y|X)$. An accurate estimate of this conditional probability is a sufficient, but certainly not a necessary, condition for minimizing the classification error. A weaker requirement that is still sufficient is to estimate the set of inputs for which $P(Y = +1|X) > 1/2$. In most cases, this requirement is much weaker than the requirement of getting good estimates of conditional probabilities. For example, if $P(Y = +1|X) = 0.49$ then our estimate of the conditional probability need be accurate to within 1%, while if $P(Y = +1|X) = 0.2$ the accuracy we need is only 30%.

This simple observation demonstrates a crucial shortcoming in the statistical interpretation of Adaboost, and undermines many of its apparent consequences, including the following:

- *Adaboost can be interpreted as a method for maximizing conditional likelihood.* If the goal is not to estimate the conditional probability, there is no reason to maximize likelihood.

- *A question of central importance is whether Adaboost is asymptotically consistent.* When evaluating probability estimators, it is standard procedure to start by verifying that the estimator is unbiased. Once the estimator is confirmed to be unbiased, the next question is the rate at which its variance decreases with the size of the sample. Again, as the learning problem in the case of classification is a weaker one, it is not clear that this is the relevant sequence of questions that a theoretician should ask.

- *Decision stumps should be used as base classifiers when the input variables are independent* This argument is based on the assumption that the goal is to estimate probabilities.

The view of AdaBoost as a method for minimizing exponential loss, though in some ways quite useful, can also lead us very much astray, as pointed out to some degree by Mease and Wyner. Taken to an extreme, this view suggests that any method for minimizing exponential loss will be equally effective, and is likely to be much better if designed with speed and this explicit goal in mind. However, this is quite false. Indeed, any real-valued classifier F which classifies the training examples perfectly, so that $y_i F(x_i) > 0$ for each training example (x_i, y_i), can be modified to minimize the exponential loss $\sum_i e^{-y_i F(x_i)}$ simply by multiplying F by an arbitrarily large positive constant. This scaling of F of course has no impact on the classifications that it makes. Thus, in the common case in which an exponential loss of zero is possible, minimization of this loss means nothing more than that the computed classifier F has a classification error of zero on the training set. The minimization of this particular loss tells us nothing more, and leaves us as open to overfitting as any other method whose only purpose is minimization of the training error.

This means that, in order to understand AdaBoost, which does indeed minimize exponential loss, we need to go well beyond this narrow view. In particular, we need to consider the *dynamics* of AdaBoost—not just *what* it is minimizing, but *how* it goes about doing it.

Like other interpretations of AdaBoost, although the statistical view has its weaknesses, it also has its strengths, as noted above. Still, to fully understand AdaBoost, particularly in the face of such deficiencies, it seems unavoidable that we consider a range of explanations and modes of understanding. Where the statistical view may be lacking, the margins explanation in particular can often shed considerable light.

Briefly, the *margin* of a labeled example with respect to a classifier is a real number that intuitively measures the confidence of the classifier in its prediction on that example. More precisely, in the notation of Mease and Wyner, the margin on labeled example (x,y) is defined to be $yF_M(x)/\sum_m \alpha_m$. Equivalently, viewing the prediction of AdaBoost's combined classifier as a weighted majority vote of the base classifiers, the margin is the weighted fraction of base classifiers voting for the correct label minus the weighted fraction voting for the incorrect label.

The margins theory (Schapire et al., 1998) provides a complete analysis of AdaBoost in two parts: First, AdaBoost's generalization error can be bounded in terms of the distribution of margins of training examples, as well as the number of training examples and the complexity of the base classifiers. And second, it can be proved that AdaBoost's dynamics have a strong tendency to increase the margins of the training examples in a manner that depends on the accuracy of the base classifiers on the distributions on which they are trained.

This theory is quite useful for understanding AdaBoost in many ways (despite a few shortcomings of its own—see, for instance, Breiman (1999) as well as the recent work of Reyzin and Schapire (2006)). For starters, the theory, in which performance depends on margins rather than the number of rounds of boosting, predicts the same lack of overfitting commonly observed in practice. The theory provides non-asymptotic bounds which, although usually too loose for practical purposes, nevertheless illuminate qualitatively how the generalization error depends on the number of training examples, the margins, and the accuracy and complexity of the base classifiers. Finally, the theory is concerned directly with classification accuracy, rather than the algorithm's ability to estimate conditional probabilities, which is in fact entirely irrelevant to the theory.

Moreover, some of the phenomena observed by Mease and Wyner do not appear so mysterious when viewed in terms of the margins theory. For instance, the experiments in Section 3.1 show AdaBoost overfitting with stumps but not decision trees. In terms of margins, decision trees have higher complexity, which tends to hurt generalization, but also tend to produce much larger margins, which tend to improve generalization, an effect that can easily be strong enough to compensate for the increased complexity. Moreover, according to the theory, these larger margins tend to provide immunity against overfitting, and indeed, overfitting is expected exactly in the case that we are using base classifiers producing small margins, such as decision stumps. This is just what is observed in Figure 1.

In sum, the various theories of boosting, including the margins theory and the statistical view, are all imperfect but are largely complementary, each with its strengths and weaknesses, and each providing another piece of the AdaBoost puzzle. It is when they are taken together that we have the most complete picture of the algorithm, and the best chances of understanding, generalizing and improving it.

References

Leo Breiman. Prediction games and arcing classifiers. *Neural Computation*, 11(7):1493–1517, 1999.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1/2/3), 2002.

Ayhan Demiriz, Kristin P. Bennett, and John Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46(1/2/3):225–254, 2002.

Yoav Freund and Robert E. Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 325–332, 1996.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 38(2):337–374, April 2000.

Jyrki Kivinen and Manfred K. Warmuth. Boosting as entropy projection. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 134–144, 1999.

Guy Lebanon and John Lafferty. Boosting and maximum likelihood for exponential models. In *Advances in Neural Information Processing Systems 14*, 2002.

Lev Reyzin and Robert E. Schapire. How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.