Journal of Machine Learning Research 9 (2008) 181-186 Published 2/08

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

Peter J. Bickel BICKEL@STAT.BERKELEY.EDU
Department of Statistics
University of California
Berkeley, CA 94720-3860, USA

Ya'acov Ritov YAACOV.RITOV@GMAIL.COM
Department of Statistics
The Hebrew University of Jerusalem
91905 Jerusalem, Israel

Editor: Yoav Freund

Galileo: God help us, I'm not half as sharp as those gentlemen in the philosophy department. I'm stupid, I understand absolutely nothing, so I'm compelled to fill the gaps in my knowledge... Sir, my branch of knowledge is still avid to know. The greatest problems still find us with nothing but hypotheses to go on. Yet we keep asking ourselves for proofs. Brecht (1980)

This is a "sock it to them" paper—though much less so than the previous versions one of us have seen.

The authors argue in the paper that AdaBoost (without early stopping) is one of, if not the, most successful boosting algorithms, and they present this paper as a disproof of what the , rather amorphous community of statistical practitioners, represented by Friedman, Hastie and Tibshirani have:

(i) Pointed out as remediable flaws of the original Freund-Schapire boosting algorithm;

(ii) Given as remedies.

Evidently, that community should be able to respond on its own. We in fact, agree with some of the hypotheses Mease and Wyner's limited simulations lead them to, whether these are or are not embraced by statistical practitioners. But others we find dubious and unproven. Let us stress the positive first.

1. They argue that boosting does not behave like nearest neighbor for $d > 1$. Not only do we agree with this but would conjecture even further without any proof:

2. That, for reasonable sequences of d dimensional distributions, the random classification rules induced by the stationary measures corresponding to boosting forever, should in a suitable sense as $n, d \to \infty$ concentrate near the Bayes rule. However, an example below shows that the improvement from $d = 1$ to $d = 2$ can be slight.

3. We don't believe that boosting is consistent, in the sense of section 3.10, for any d, but indeed there is no disproof for $d > 1$.

4. We agree that a sharp explanation why, for classification, boosting *may* not overfit—that is, continues to reduce the probability of misclassification long past the point where all training sample observations are correctly classified has not been provided in the statistical (or the machine learning) literature.

5. We agree that using more complex basis functions may actually improve performance. This was analyzed theoretically for L_2 boosting by Bühlmann and Yu (2003).

6. We agree that there is a need for a convincing argument for basing an early-stopping algorithm of a classifier on a loss function that is not the classification loss. *A-priori* we do not expect that stopping on any criterion, other than minimum classification error will work in general, even if the classifier itself is based on minimizing this indirectly relevant criterion. However, it certainly can be that a good early stopping algorithm will be based on estimate of the loss with respect to something other than from the classification error. It was proven to be so, for example, with L_2-boost.

Since we have never been persuaded on theoretical grounds of the superiority of other "boosts", logit or L_2, over AdaBoost , we leave this battle to others.

Where we really part company with the authors of this ”against the heretics” paper is on the issue of the desirability of early stopping.

> *Galileo tries to explain his young student, Andrea, the structure of the Copernican system, to make it so simple that Andrea will be able to explain it to his mother. He rotates him on a chair, and tells him that an apple in the center is the earth. However, Andrea is smart enough to understand that so far and not more can be deduced from examples.*
> Andrea: Examples always work if you're clever. Only I can't lug my mother round in a chair like you did me. So you see it's a rotten example really. And suppose your apple is the earth like you say? Nothing follows. Brecht (1980)

Everybody can produce examples. The authors gave two examples. Other commentators will bring their own. Here are ours. We consider $X \in \mathbb{R}^2$ uniform on 5 concentric circles. The classes were randomly assigned according to $P(Y = 1|x) = \text{logit}(4\text{sgn }(\xi)\sqrt{(\xi)})$ where $\xi = \|x\| - 2|x_1|$ and $\text{logit}(\zeta) = e^{\zeta}/(1+e^{\zeta})$. The training sample includes 500 i.i.d. observations. 200 more observations were used for early stopping. I.e., stopping when the mean empirical classification error on these 200 observations was minimal. Finally another set of 1500 observations was used to evaluate the mean classification error of the AdaBoost procedure as a function of the number of observations. See Figure 1 for a plot of typical set of Xs and $P(Y = 1|X))$. The weak classifier we used was a standard classification tree with 8-terminal nodes. The simulations were run with slightly different set-ups a few tens of times. There was no case that contradicted the results of the single experiment we will present next (but see later). Our example is small. Since the goal of the discussed paper is to suggest policy on the basis of two examples, even one (we believe) reasonable counterexample should give some pause.

In Figure 2(a) we show the classification error as a function of the number of iterations. The horizontal lines are, from the top down: The risk of the closest neighbor, the risk of AdaBoost with early stopping, and the Bayes Risk. It is clear in this example that AdaBoost starts quite nicely. The primitive early stopping technique we employed is enough to give a decent performance. However, the performance of AdaBoost starts to degenerate after some tens of iterations. After 800 hundred

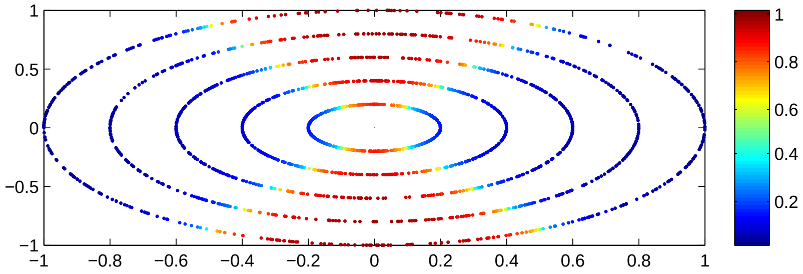


Figure 1: $P(Y=1|X)$ as function of location.

iterations it is only slightly better than the nearest neighbor classifier. Within the context of this example the "does not over-fit" property, can be understood at best as "it is simply a slow algorithm".

Figure 2(b) shows the root mean square error of AdaBoost implicit estimate of the probability as function of the number of iterations. It is clear that it degenerate much faster than AdaBoost performance. However, the implicit probability estimate is fair, as long as the classifier is in its prime. So, the authors' doubt whether boosting estimate probabilities cannot be based on this example.

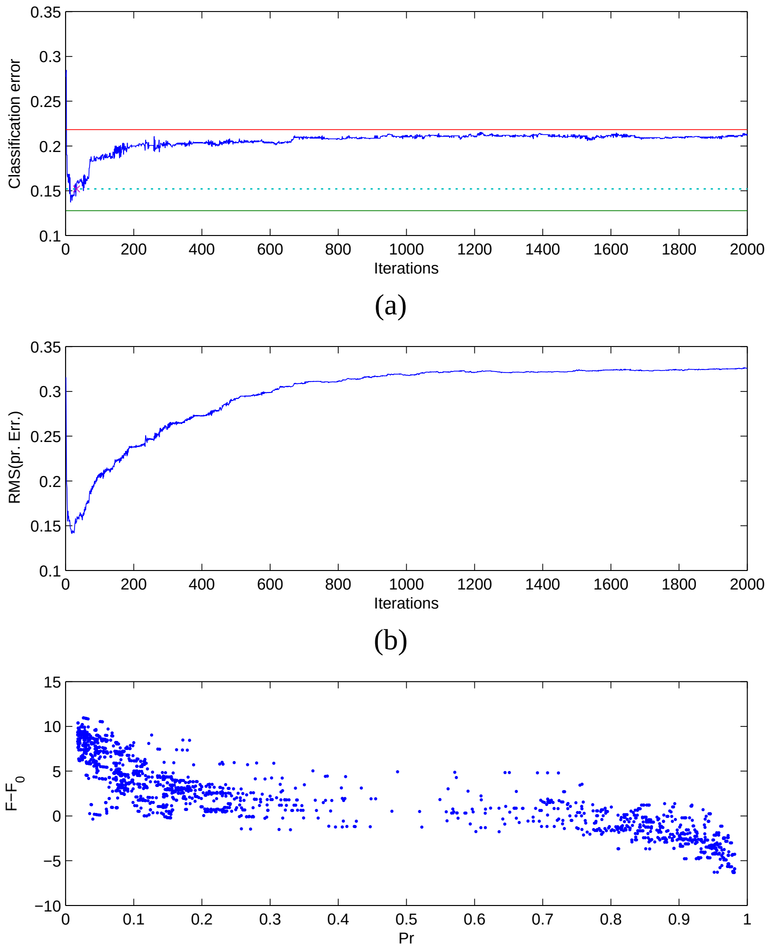
However, we should mention that the apparent failure of the boosting algorithm to estimate probabilities is somewhat misleading. In Figure 2(c) we plot $F_m - F^0$, where F^0 is the ideal value, as a function of $P(Y=1|x)$ after $m=200$ iterations. $P(Y=1|x)$ is used here as proxy for the distance of the point from the $P(Y=1|x)=0.5$ boundary. What can be seen is that most of the error in the estimation of the $P(Y=1|x)$ comes from the easy to classify points. So, the boosting algorithm fails to estimate the probability where it does not really matter.

There are facts in life. One cannot invoke the church teaching or Aristotle's books in face of empirical facts. Galileo get annoyed by the insistence of the philosopher and the mathematician on using irrelevant arguments.

Galileo: My reasons! When a single glance at the stars themselves and my own notes make the phenomenon evident? Sir, your disputation becoming absurd. Brecht (1980)

The authors used a smart device to present the discrepancy between the one nearest neighbor classifier and AdaBoost. Namely, they compare their performance under a null hypothesis, where $P(Y=1|x)$ does not depend on x (this is true for Section 3.9 but not for 4.9, where for some unknown reasons something else was done). We did the same. The distribution of X and the sample sizes are as above, while $P(Y=1|x) \equiv 0.2$. The results are presented in Figure 3. The graphs are similar to Figure 2(a-b), in description and in essence. Boosting without early stopping is not like 1-NN, but it is not much better, at least for this tiny example.

We did observe a further interesting phenomenon when we added irrelevant explanatory variables. That is, 8 independent variables $X_3, \ldots, X_{10}$ were added, while the distribution of (X_1, X_2, Y) remains as above. The result was surprising. Adding these irrelevant variables improved the performance of "boosting-for-ever" to the level of the early stopping algorithm. This gives some credence to our conjecture 2.

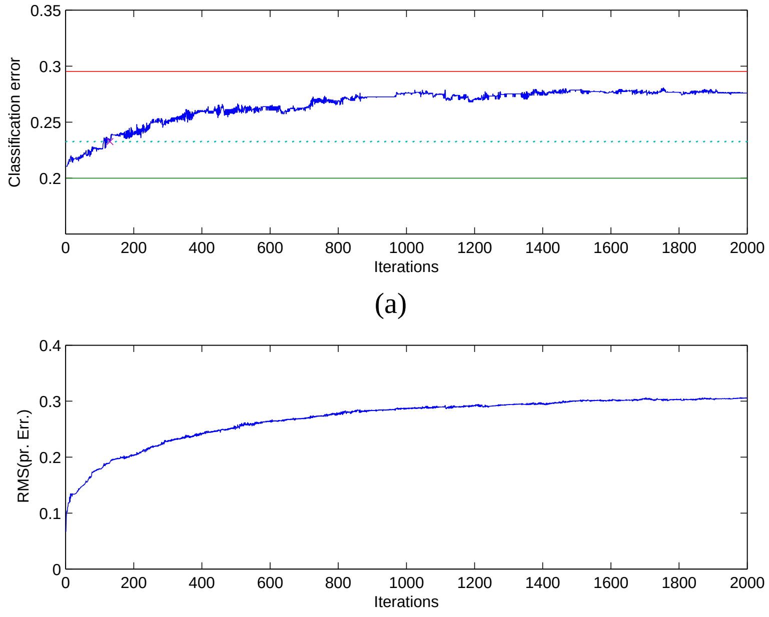


(a)

(b)

(c)

Figure 2: AdaBoost, with and without early stopping.



(a)

(b)

Figure 3: AdaBoost under the null hypothesis.

But, this is not the end of the story. We changed the the distribution a little. We left $P(Y|X_1,\ldots,X_{10}) = P(Y|X_1,X_2)$ as above. $X_3,\ldots,X_{10}$ are still independent of each other and independent of (X_1,X_2). However the distribution of (X_1,X_2) was changed to be discrete with 24 well isolated atoms. One can conceive the distribution as that of 24 columns with $500/24$ on the average observations. The Ys are i.i.d. given the column. AdaBoost with early stopping handled this situation very well and stopped after very few (e.g., 2) iterations. It essentially isolated the columns and left them intake. Boosting-for-ever stabilized nicely after very few iterations. However, as could be expected, it did break the columns. The result was that the misclassification error of boosting-for-ever was almost in the middle between the early stopping algorithm, and the strictly inferior 1-NN estimator.

A further point. The phenomenon of not overfitting for a long time is certainly interesting to investigate, but why this should be a virtue of a procedure is unclear, since it merely increases computation time at an often (perhaps usually) negligible improvement over stopping early.

AdaBoost is a mystery, but we, the weak, can solve only one toy problem at a time.

Galileo: Why try to be so clever now, that we at last have a chance of being less stupid? Brecht (1980)

AdaBoost was crowned by Leo Breiman as the best off-the-shelf classifier. It has some mysterious properties, particularly, sometimes continuing to improve off-sample performance even after completely collapsing on the data. It behaves better, sometimes, with 8-nodes trees, some other times with 256-node trees, and other times, mainly when examined by some statisticians, stumps are superior. It presents some mathematical challenges, which should be carefully investigated. However, many examples appearing in the literature are either very artificial, or the investigators don't have a gold standard like the Bayes risk. We hope that this technique will grow out of its status of something like an art, to a scientifically justified method. But this is only a hope.

Acknowledgments

This work was supported in part by NSF grant DMS-0605236 and an ISF grant.

References

Bertolt Brecht. *Life of Galileo*. Arcade Publishing, New York, 1980. Translated by John Willet.

Peter Bühlmann and Bin Yu. Boosting with the l2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.