Journal of Machine Learning Research 9 (2008) 195-201 Published 2/08

Evidence Contrary to the Statistical View of Boosting: A Rejoinder to Responses

David Mease MEASE_D@COB.SJSU.EDU
Department of Marketing and Decision Sciences
College of Business, San Jose State University
San Jose, CA 95192-0069, USA

Abraham Wyner AJW@WHARTON.UPENN.EDU
Department of Statistics
Wharton School, University of Pennsylvania
Philadelphia, PA, 19104-6340, USA

Editor: Yoav Freund

1. Introduction

We thank the discussants for their comments on our paper. We also thank the editors for arranging the discussions. Many interesting points have been raised by the discussants. We can not respond to everything, but we do include a section addressing the main points of each discussant. Following these, we provide a final section in which we give some general concluding remarks.

Many of the discussants comment on the overfitting of boosting. Different authors will have different ideas of what the term overfitting means in the context of boosting, but for clarification throughout this rejoinder we will define overfitting as a positive slope for a specified loss metric as a function of the iterations. Specifically, the loss metric we focus on is misclassification error, although we understand that some of the discussants are concerned about other loss functions which quantify probability estimation accuracy rather than classification accuracy. The importance of focusing on misclassification error is underscored in the discussion by Freund and Schapire who remind us that AdaBoost is an algorithm for carrying out classification, not probability estimation.

2. Rejoinder for Kristin P. Bennett

Bennett provides a useful perspective on the situation by studying the convergence of boosting algorithms from the optimization point of view. We agree that this aspect of the problem is too often overlooked by researchers.

In studying the convergence of the algorithms, Bennett touches on a number of important considerations. For example, she mentions cycling in the context of stumps and notes that the cycling results in boosting using only a small number of unique trees. The number of unique trees is rarely noted by researchers in empirical investigations. Bennett's studies also lead her to concur that boosting algorithms sometimes benefit from a bagging type of self-averaging during what she calls the "overtraining" stage. ("Overtraining" as defined by Bennett should not be confused with "over*fitting*" as we have defined it). Another point on which we strongly agree with Bennett is that boosting's resistance to overfitting can occur for different reasons in different contexts. We believe that this is one reason why researchers have difficulty in coming to agreement regarding various explanations for boosting.

In addition to studying convergence, Bennett also looks at the margin of the classification rules. We mentioned margin theory in our paper only briefly since our focus was on the statistical view of boosting, and margin theory is generally separate from this view. It is our hope, however, that by finding holes in the accepted statistical view we are encouraging researchers to approach the problem from different perspectives to help explain the phenomena left unexplained by the statistical view. Unfortunately, in this case Bennett points out that examination of the margin still fails to explain the results of the experiment in Section 3.2.

We believe that studying boosting as an algorithm (rather than as a statistical model) in the way Bennett has done can be quite helpful in understanding some of its remaining mysteries. We are glad that our examples have inspired this type of investigation.

3. Rejoinder for Andreas Buja and Werner Stuetzle

One of the main points argued by Buja and Stuetzle is that most of the current literature on boosting does not explain its variance reduction. They argue that a complete view of boosting should explain both its ability to reduce bias and variance. We certainly agree with this point. In fact, some of the examples in our paper such as those in Sections 3.4 and 4.4 illustrate this quite well. It is interesting that Buja and Stuetzle site references from the early research on boosting which argue in this same direction. It is a shame that more attention has not been given to boosting's ability to reduce variance in addition to bias in more recent research. We hope that our paper helps to rejuvenate research on this aspect of boosting.

Buja and Stuetzle go on to argue that there is often a need for more than a black box classifier which produces small misclassification error. Some applications call for interpretable and diagnostic models and/or conditional class probability estimates. This is certainly true, and we agree that research that extends boosting in these directions is quite welcome. However, in carrying out this research it is important to be honest about situations in which the theory does not explain the performance of traditional boosting algorithms for classification. One of our purposes in writing this paper was to promote this honesty.

4. Rejoinder for Yoav Freund and Robert E. Schapire

Freund and Schapire focus their discussion on margin theory. It is quite interesting that margin theory has not been embraced much at all by the statistical community. In fact, Freund and Schapire's paper "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods" appeared less than two years before the Friedman, Hastie and Tibshirani paper "Additive Logistic Regression: A Statistical View Of Boosting" in the same journal (*Annals of Statistics*). Despite this, the statistics community has largely ignored it in favor of the more familiar theory in the latter paper.

Freund and Schapire make a case for margin theory by arguing that this theory explains the results of the experiment in Section 3.1 while obviously the statistical theory does not. However, it should be noted that Bennett believes margin theory still does not explain the results for Section 3.2. Despite this, we still believe that margin theory is worth pursuing, especially given the large number of inconsistencies between the statistical view and the reality of the simulations presented in our paper. As we have mentioned, we wrote the paper with the goal of promoting alternative explanations for boosting other than the statistical view, which leaves much unexplained. Margin theory is one such alternative explanation.

Additionally, Freund and Schapire make a number of other points on which we agree and wish to underscore. They argue that in the statistics research on boosting too much importance is placed on class probability estimation over class estimation, the use of stumps over larger trees and theoretical questions of consistency and asymptotic variance over more relevant theoretical questions.

5. Rejoinder for Jerome Friedman, Trevor Hastie and Robert Tibshirani

Friedman, Hastie and Tibshirani argue that many of the ten statements made in our paper should not be ascribed to the statistical view as laid out in Friedman et al. (2000). This is understandable, and in the case of the similarity with nearest neighbor algorithms, for example, we have even noted this in our original paper. However, other statements could be argued to follow fairly directly, whether that is the intent of the authors or the fault of the reader. At the very least, we believe it is fair to say the statistical view offers very little to help explain the non-intuitive nature of the results in our paper.

Of all ten statements, the most direct relationship to the work of Friedman, Hastie and Tibshirani is the idea in Sections 3.1 and 4.1 that stumps should be used for additive Bayes rules. We believe the authors would agree that this is the strongest connection to their work, which is why they focused the majority of their discussion on the experiment in Section 3.1. We also think it is refreshing that they admit that they “are not sure why” the results of this experiment are such, but they do produce some graphs to try to understand this better. Some of their graphs show the performance of the probability estimates. It is argued in the discussion by Freund and Schapire that probability estimation for boosting has very little to do with boosting’s classification performance. We agree and for this reason we will not comment on the graphs for probability estimation. However, the graphs showing misclassification error are of interest and we discuss these below.

Friedman, Hastie and Tibshirani’s Figure 1 shows that using shrinkage in our original experiment causes stumps to “win handily”. We note, however, that the shrinkage causes overfitting, so it also becomes necessary to stop the boosting process before 600 iterations to realize any advantage over the 8-node trees. In practice the optimal stopping time is not known, and a fair comparison would require incorporating the uncertainty in the estimation of this value.

The right panel of their Figure 2 shows that with a larger training sample size of $n = 2000$ the overfitting caused by shrinkage is not as severe and the stumps maintain their advantage over the full 1000 iterations. However, by 1000 iterations the overfitting for the stumps has caused the gap with the 8-node trees to close considerably and extrapolation would suggest that additional iterations would result in this gap becoming even smaller or disappearing altogether. Meanwhile, the 8-node trees again show no signs of overfitting.

We believe the more interesting research question with regard to Friedman, Hastie and Tibshirani’s Figures 1 and 2 is not which algorithm performs best for a certain stopping time and sample size, but rather why all algorithms display overfitting with regard to misclassification error except the 8-node trees without shrinkage. The authors state that the larger trees can have a bagging effect, and we certainly agree with this. We feel that understanding this effect better (as well as understanding why shrinkage can destroy this effect) is essential to gaining a better understanding of boosting.

The left panel of Friedman, Hastie and Tibshirani's Figure 3 shows the effect of using Bernoulli loss rather than exponential loss. With Bernoulli loss all algorithms show overfitting. This is another curiosity not explained by the statistical view of boosting. In fact, the paper by Friedman et al. (2000) seems to suggest the opposite should be expected.

In their final paragraph Friedman, Hastie and Tibshirani welcome "*constructive* counter-arguments and alternative explanations for the success of boosting methods." We will make a couple of comments here in response to this. First, we believe that explanations such as the variance reducing bagging effect of boosting fall into this category. Our paper is certainly not the first to suggest this notion, nor do we offer a theoretical explanation for the phenomenon, but our paper does stress the importance of not overlooking this effect and thus we hope promotes constructive research on this topic. Secondly, we feel that researchers currently embrace the statistical view too strongly, and for this reason it is difficult for researchers to offer any alternative explanations without first tearing down the current beliefs to some degree. We base this statement on our own experience. For instance, an early version of Mease et al. (2007) was rejected for publication by a different outlet. That paper offers a method for estimating probabilities using AdaBoost. Two of the three referees rejected the paper arguing that in order to estimate probabilities using boosting it would be sufficient to use LogitBoost in place of AdaBoost.

6. Rejoinder for P. J. Bickel and Ya'acov Ritov

Bickel and Ritov begin by identifying some points we made in our paper with which they agree. The amount of agreement is substantial. The disagreement is focused largely on what is generally regarded to be the most mysterious property of the AdaBoost algorithm: its ability to reduce the (test) error rate long after the training data has been fit without error. Other classifiers such as CART, neural nets, LDA and logistic regression perform optimally only with some sort of appropriate early stopping to prevent over parameterization and overfitting of the data. It is understandable that Bickel and Ritov's negative remarks focus on this particular issue, as it is the feature of AdaBoost most at odds with the statistical view.

The example provided by Bickel and Ritov is a model for which early stopping is essential to achieve optimal classification performance. If AdaBoost is not stopped after 10 or 20 iterations in their example, it will overfit the data and the generalization error will increase steadily. Their example is not the first of its kind, but rather is typical of the simulations used to provide empirical support for the statistical view. We addressed this point directly in our paper:

> *Such examples in which overtting is observed often deal with extremely low-dimensional cases such as $d = 2$ or even $d = 1$. By experimenting with the simulation code provided along with this paper, one can confirm that in general AdaBoost is much more likely to suffer from overtting in trivial low-dimensional examples as opposed to high-dimensional situations where it is more often used.*

The example provided by Bickel and Ritov is of this same spirit. It is a $d = 2$ dimensional model that lives on $d = 1$ dimensional circular manifolds. Since it is well known that AdaBoost will converge to the nearest neighbor classifier in one dimension, the results for this simulation are not unexpected. Along these same lines, Bickel and Ritov provide further evidence for our claim that overfitting is largely a symptom of trivial low dimensional examples by observing that when they add 8 additional dimensions with no signal the overfitting disappears. The authors refer to this phenomenon as “surprising”, but we would argue that this should also be expected.

Bickel and Ritov note interesting changes when they discretize some of the predictors. We did not discuss this in our paper, but the difference in behavior for discrete data and continuous data is tremendous. We learned first hand about this in Mease et al. (2007) when we artificially introduced discreteness by adding replicates. Because of this difference, we focused our current paper on the case of continuous data. The case of discrete data relates more to what Friedman et al. (2000) called the *population version* of boosting. The statistical view of boosting is largely based on this population version of boosting, and thus the statistical view becomes more relevant for discrete data.

Bickel and Ritov also bring up probability estimation in their discussion. Probability estimation is a popular topic among statisticians with regard to boosting, but again we will not go into any detail here because we agree with Freund and Schapire that probability estimation is not the central topic, but rather classification. It is curious, however, that Bickel and Ritov state that “most of the error in the estimation” of the probabilities “comes from the easy to classify points”. The truth of this statement depends on how one measures the error in probability estimates. For example, if one computes the RMSE between the true probabilities and the estimated probabilities for the data in Bickel and Ritov’s Figure 2(c), it can be observed that the RMSE is actually highest near the points with a true probability of 1/2.

In conclusion, Bickel and Ritov have presented a couple of low dimensional examples in which overfitting occurs unless early stopping is applied, while our paper presents higher dimensional examples where this is not the case. Since “everybody can produce examples” as Bickel and Ritov state, one may wonder which set of examples is more useful to study for the purpose of understanding boosting. We make the argument for our examples based on the fact that higher dimensional examples are more common of a use case for boosting, and that our examples are more similar to the many examples in which overfitting does not occur that first led practitioners to be attracted to boosting originally. We feel that understanding such examples is most useful for understanding boosting.

7. Rejoinder for Peter Buhlmann and Bin Yu

Of the six discussions, Buhlmann and Yu seem to be the least accepting of the empirical findings in our paper. They state that the “main reason” why we obtain contrary evidence in many cases is the functional form of the simulation model, and they propose considering a model that is additive on the logit scale. The model they propose turns out to be a special case of our model in Section 5 with $k = 8$. As stated in the paper, we had already considered the results for this specific model carefully using various values of k, and those results were consistent qualitatively with the other simulations in the paper. Thus, we were surprised to read that Buhlmann and Yu based any disagreement on results from this model. On closer inspection, we learned that the discrepancy is most likely due to sampling error. Buhlmann and Yu considered only a single repetition. The plot in the first panel of their Figure 3 shows that for that single repetition, the 8-node trees begin to overfit near 400 iterations and the stumps have a lower misclassification error throughout the 1000 iterations. However, if the results are averaged over many repetitions, one can learn that this behavior is not true in aggregate. The first plot in Figure 1 in this rejoinder shows the misclassification error for the Buhlmann and Yu model averaged over 100 repetitions. The stumps overfit while the 8-node trees do not, and the 8-node trees achieve a lower misclassification error than the stumps at 1000 iterations. Thus, the results for the Buhlmann and Yu model are consistent with the results for the other simulations in our paper.

It should be noted that with the small sample size of $n = 100$ chosen by Buhlmann and Yu, there is indeed often a problem with floating point overflow errors in R as they mention. They dealt with this issue by making a modification to our code (which is an alternative explanation for the discrepancies in the results). We avoided this issue when making the first plot in Figure 1 in this rejoinder by considering 100 randomly chosen repetitions for which the floating point overflow error did not occur. For readers who are uncomfortable with this, there are a couple of other ways of handling the problem. First, if a larger sample size is chosen then the problem goes away completely. For example, the second plot in Figure 1 in this rejoinder considers our original sample size of $n = 200$. For this sample size there were no problems with overflow errors in 100 repetitions. Again, one can see that the 8-node trees lead to a lower misclassification error at 1000 iterations and do not overfit, while the stumps do overfit. A second way of dealing with the overflow errors for the sample size of $n = 100$ is to consider only the first 500 iterations. When we ran only 500 iterations we did not observe any overflow errors in 100 repetitions, and the results were consistent with those in the first plot in Figure 1 in this rejoinder.

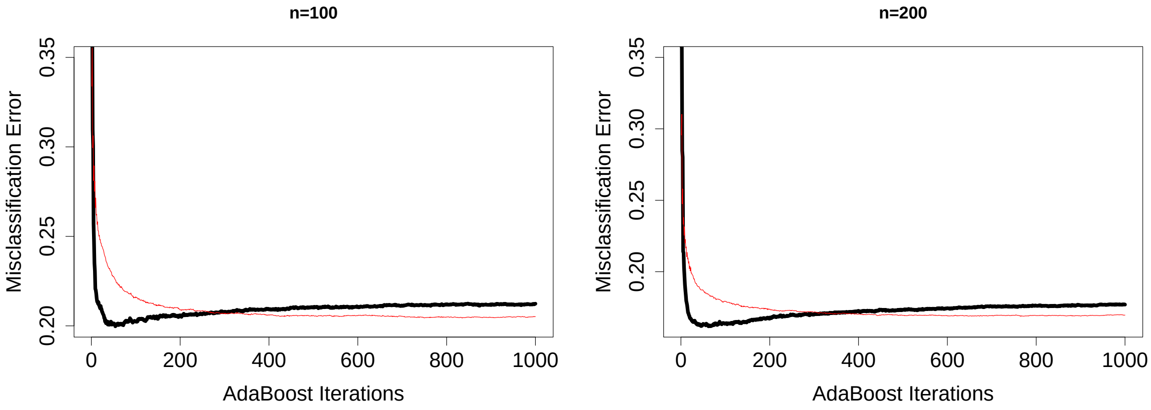

Figure 1: Comparison of AdaBoost with Stumps (Black, Thick) and 8-Node Trees (Red, Thin) for the Buhlmann and Yu Model Using 100 Repetitions and Sample Sizes of $n = 100$ and $n = 200$

Thus, concerning this simulation model proposed by Buhlmann and Yu, we conclude that the results are largely consistent with the other results in our paper. However, Buhlmann and Yu also discuss our Figures 3 and 4 independently of their simulation model. We address this below.

With regard to LogitBoost in our Figure 3, Buhlmann and Yu mention that with early stopping LogitBoost does almost as well as AdaBoost. This assumes one can estimate the optimal stopping time well, when in practice the stopping time estimation using LogitBoost can be difficult. In fact, we mentioned in our paper that the authors of the particular LogitBoost code we used reported that the stopping estimation was not effective for their purposes. Conversely AdaBoost performs fine in our Figure 3 without early stopping. Also, considering that the creators of LogitBoost (Friedman, Hastie and Tibshirani) mentioned in their discussion that they “have moved on” from LogitBoost, there seems to be little left to be said in its defense.

With regard to our Figure 4, Buhlmann and Yu seem unimpressed by this example which shows how AdaBoost can reduce variance. In fact, despite its excellent performance they lament that the classifier "gives the wrong impression." This is an interesting statement and illustrates the desire on the part of the statistical community to view AdaBoost as an interpretable model rather than a black box classifier. Certainly a model can give one the wrong impression but a classifier arguably only classifies well or does not classify well.

In conclusion, Buhlmann and Yu are considerably opposed to our most important claims: 1) that large trees generally work better than stumps and 2) that overfitting is usually not a problem and 3) that early stopping initiatives are often not only unnecessary but also counterproductive. With regard to classification we have shown that their simulation does not provide strong evidence to support their case. However, when the goal is the more difficult problem of conditional probability estimation, we will not offer any disagreement. We have not focused our analysis on this problem, and we do not feel that boosting algorithms should be regarded as "state-of-the-art" probability estimators, since they are usually outperformed by competitors like Random Forests or even logistic regression. Rather, it is the statistical view that asserts that AdaBoost works, erroneously in our opinion, because it estimates probabilities. Indeed, the efforts of Buhlmann and Yu to understand and improve the performance of boosting for probability estimation is productive and worthwhile. Where we part company is in name only. We question whether it is logical to continue to call the algorithms boosting algorithms since there is a considerable disconnect from the original AdaBoost algorithm for classification.

8. Conclusion

We believe the discussions provided by the six sets of authors have been extremely valuable. We are encouraged by the amount of discussion of two main ideas which we feel will lead to a better understanding of boosting.

First, the discussions have helped to clarify that (out of sample) minimization of the surrogate loss function (and equivalently probability estimation) is often a very different problem from (out of sample) minimization of misclassification error. The original AdaBoost algorithm was intended for the latter but much research in the statistical community has focused on the former.

Secondly, with regard to minimization of misclassification error, the idea that boosting is reducing variance has been acknowledged in the discussions. The extent to which this phenomenon is essential for boosting's success and of interest as a research topic can be debated, but we are encouraged to see it acknowledged by such a prominent group of researchers.

References

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–374, 2000.

D. Mease, A. Wyner, and A. Buja. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8:409–439, 2007.