Training-Time Optimization of a Budgeted Booster

Yi Huang, Brian Powers, Lev Reyzin
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607
{yhuang,bpower6,lreyzin}@math.uic.edu

Abstract
We consider the problem of feature efficient prediction – a setting where features have costs, and the learner is limited by a budget constraint on the total cost of the features it can examine in test time. We focus on solving this problem with boosting by optimizing the choice of base learners in the training phase and stopping the boosting process when the learner’s budget runs out. We experimentally show that our method improves upon the boosting approach AdaBoostRS (Reyzin, 2011) and in many cases also outperforms the recent algorithm SpeedBoost (Grubb and Bagnell, 2012). We also experimentally show that our method also outperforms pruned decision trees, a natural budgeted classifier.

Introduction
The problem of budgeted learning centers on questions around resource constraints imposed on a traditional supervised learning algorithm. Here, we focus on the setting where a learner has ample resources during training time, but is constrained by resources in predicting on new examples. In particular, we assume that accessing the features of new examples is costly (with each feature having its own cost to access it), and predictions must be made without running over a given budget. This budget may or may not be known to the learner. Learners that adhere to such budget constraints are sometimes called feature-efficient.

A classic motivation for this problem is the medical testing setting, where features correspond to the results of tests that are often costly or even dangerous to perform. Diagnoses often need to be made on incomplete information, and doctors must order tests thoughtfully in order to stay within whatever budgets the world imposes.

Here, we focus on boosting methods, in particular AdaBoost, to make them feature-efficient predictors. This line of work was started by Reyzin (2011), who introduced the algorithm AdaBoostRS, a feature-efficient version of AdaBoost. While AdaBoostRS provably converges to the behavior of AdaBoost as the feature budget increased, it only considers feature costs and budget at test time. Reyzin left open the problem of whether optimizing during training can improve performance. Here, we answer this question with a resounding yes, giving algorithms that clearly outperform AdaBoostRS, especially when costs vary and budget limits are small.

Our approach relies mainly on two observations. The first is that when all features have equal costs, stopping the training of AdaBoost early, once the budget runs out, will outperform AdaBoostRS. Second, when features have different costs, which is the setting that chiefly concerned Reyzin, one can still run AdaBoost, but choose weak learners as to better trade-off their cost vs. contribution to the performance of the ensemble. Combining these simple ideas yields our approach.

Past work
Research on this problem goes back at least to Wald (1947), who considered the problem of running a clinical trial sequentially, only testing future patients if the validity of the hypothesis in question is still sufficiently uncertain. This question belongs to the area of sequential analysis (Chernoff, 1972).

Ben-David and Dichterman (1993) examined the theory behind learning using random partial information from examples and discussed conditions for learning in their model. Greiner et al. (2002) also considered the problem of feature-efficient prediction, where a classifier must choose which features to examine before predicting. They showed that a variant of PAC-learnability is still possible even without access to the full feature set.

In related settings, Globerson and Roweis (2006) looked at building robust predictors that are resilient to corrupted or missing features. Cesa-Bianchi et al. (2010) studied how to efficiently learn a linear predictor in the setting where the learner can access only a few features per example. He et al. (2012) trained an MDP for this task, casting it as dynamic feature selection – their model is a variant of ours, except that they attempt to jointly optimize feature costs and errors, whereas our model has a strict bound on the budget. Finally, Xu, Weinberger, and Chapelle (2012) tackled a related a feature-efficient regression problem by training CART decision trees with feature costs incorporated as part of the impurity function.

In the area of boosting, Pelossof et al. (2010) analyzed how to speed up margin-based learning algorithms by stopping evaluation when the outcome is close to certain. Sun
and Zhou (2013) also considered how to order base learner evaluations so as to save prediction time.

However, our main motivation is the work of Reyzin (2011), who tackled the feature-efficient learning problem using ensemble methods. He showed that sampling from a weights distribution of an ensemble yields a budgeted learner with similar properties to the original ensemble, and he tested this idea experimentally on AdaBoost. The goal of this paper is to improve on Reyzin’s approach by incorporating the feature budget into the training phase.

We also compare to the recent work of Grubb and Bagwell (2012), who also focused on this setting. Their algorithm, SpeedBoost, works by sequentially choosing weak learners and voting weight $\alpha$ as to greedily optimize the improvement of a loss function (e.g., exponential loss) per unit cost, until the budget runs out.

**AdaBoost and early stopping**

Our goal in this paper is to produce an accurate classifier given a budget $B$ and a set of $m$ training examples, each with $n$ features, and each feature with a cost via cost function $C : [n] \rightarrow \mathbb{R}^+$. Reyzin’s AdaBoostRS (2011) takes the approach of ignoring feature cost during training, and then randomly selecting hypotheses from the ensemble produced by AdaBoost until the budget is reached. Here we look at a different approach – to optimize the cost efficiency of boosting during training, so the ensemble classifier that results is both relatively accurate and affordable.

One straightforward approach is to run AdaBoost, paying for the features of the weak learners chosen every round, bookkeeping expenditures and the features used, until we cannot afford to continue. In this case, we are simply stopping AdaBoost early. We call this algorithm the “basic” AdaBoostBT for Budgeted Training. Surprisingly, this albeit simple methodology produces results that are significantly better than AdaBoostRS for both features with a uniform cost and features with random cost across a plethora of datasets.

We note that, in AdaBoost, since training error is upper bounded by

$$\hat{\Pr}[H(x) \neq y] \leq \prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} \sqrt{1 - \gamma_t^2},$$

at each round of boosting one typically greedily chooses the base learner that minimizes the quantity, which is equivalent to choosing the weak learner that maximizes $\gamma_t$. This is done in order to bound the generalization error, which was shown by Freund and Schapire (1997) to be bounded by

$$\Pr[H(x) \neq y] \leq \hat{\Pr}[H(x) \neq y] + O \left( \sqrt{\frac{T \log d}{m}} \right).$$

In these bounds $\hat{\Pr}$ refers to the probability with respect to the training sample.

Hence, one can simply choose $h_t$ in step 5 of AdaBoostBT according to this rule, which amounts to stopping AdaBoost early if its budget runs out. As we show in Section 4, this already yields an improvement over AdaBoostRS. This is unsurprising as, especially when the budget or number of allowed rounds is low, AdaBoost aggressively drives down the training error (and therefore generalization error), which AdaBoostRS doesn’t do as aggressively. A similar observation will explain why the methods we will introduce presently also outperform AdaBoostRS.

However, this approach turns out to be suboptimal when costs are not uniform. Namely, it may sometimes be better to choose a worse-performing hypothesis if its cost is lower. Doing so may hurt the algorithm on that current round, but allow it to afford to boost for longer, more than compensating for the locally suboptimal choice.

**A better trade-off**

Here we focus on the problem of choosing a weak learner when feature costs vary. Clearly, higher values of $\gamma_t$ are still preferable, but so are lower feature costs. Both contribute to minimizing the quantity $\prod_{t=1}^{T} Z_t$, which upper bound the training error in different ways. High $\gamma$s make the product small-term-wise. Lower costs, on the other hand, allow for more terms in the product. The goal is to strike the proper balance between the two.

One problem is that it is difficult to know exactly how many future rounds of boosting can be afforded under most

---

Algorithm 1 AdaBoostBT($S, B, C$), where: $S \subset X \times \{-1, +1\}$, $B > 0$, $C : [n \times m] \rightarrow \mathbb{R}^+$

1: given: $(x_1, y_1), \ldots, (x_m, y_m) \in S$
2: initialize $D_1(i) = \frac{1}{m}$, $B_1 = B$
3: for $t = 1, \ldots, T$ do
4: train base learner using distribution $D_t$, get $h_t \in H : X \rightarrow \{-1, +1\}$
5: if the total cost of the unpaid features of $h_t$ exceeds $B_t$ then
6: set $T = t - 1$ and end for
7: else set $B_{t+1}$ as $B_t$ minus the total cost of the unpaid features of $h_t$, marking them as paid
8: set $\alpha_t = \frac{1}{Z_t} \ln \frac{1 + \gamma_t}{1 - \gamma_t}$, where $\gamma_t = \sum_i D_t(i) y_i h_t(x_i)$
9: update $D_{t+1}(i) = D_t(i) \exp(\alpha_t y_i h_t(x_i)) / Z_t$, where $Z_t$ is the normalization factor
10: end for
11: output the final classifier $H(x) = \text{sign} \left( \sum_{t=1}^{T} \alpha_t h_t(x) \right)$
strategies. Hence, we can make the assumption that for a base learner that costs \(c\), we could afford \(B_t/c\) additional rounds of boosting under the simplified assumption that all future rounds will incur the same cost and achieve the same \(\gamma_t\) as in the current round. In this case, minimizing the quantity \(\prod_{t=1}^{T} Z_t\) is equivalent to minimizing the quantity
\[
h_t = \arg\min_{h \in \mathcal{H}} \left( (1 - \gamma_t(h)^2)^\frac{1}{m_{t+1}} \right),
\]
where
\[
\gamma_t(h) = \sum_i D_t(i)y_i h(x_i),
\]
and \(c(h)\) is the cost of the features used by \(h\). This is our first algorithm criteria for modifying base learner selection in step of AdaBoostBT (boxed for emphasis). We call this algorithm AdaBoostBT\_Greedy.

There is a potential pitfall with this approach: if we mark every used feature down to cost 0 (since we don’t re-pay for features), then the optimization will collapse since every base learner with cost 0 will be favored over all other base learners no matter how uninformative it is. We can obviate this problem by considering the original cost during the selection, but not paying for used features again while updating \(B_t\), as done in our Algorithm.

As optimizing according to Equation 1 makes a very aggressive assumption of future costs, we consider a smoother optimization for our second approach. If in round \(t\) we were to select \(h_t\) with cost \(c\), the average cost per round thus far is clearly
\[
\frac{(B - B_t) + c^t}{t}.
\]
Our second approach uses this average cost to estimate the number of additional rounds we are going to run. Specifically, in step of AdaBoostBT, we select a base learner according to
\[
h_t = \arg\min_{h \in \mathcal{H}} \left( (1 - \gamma_t(h)^2)^\frac{1}{m - m_t} \right),
\]
Selecting according to Equation 2 is less aggressive, because as more of the budget is used, current costs matter less and less. Hence, we call this second algorithm AdaBoostBT\_Smoothed.

Additional theoretical justification

While the theory behind optimizing the bound of \(\prod_{t=1}^{T} Z_t\) on the training error is clear, we can borrow from the theory of margin bounds (Schapire et al., 1998) to understand why this optimization yields improved results for generalization error.

One might be concerned that in finding low cost hypotheses, we will be building too complex a classifier, which will not generalize well. In particular, this is the behavior that the Freund and Schapire (1997) generalization bound would predict. Fortunately, the margin bound theory can be used to alleviate these concerns.

The following bound, known as the margin bound, bounds the probability of error as a sum of two terms.
\[
P_D[yf(x) \leq 0] \leq P_S[yf(x) \leq \theta] + O\left( \frac{d}{m\theta^2} \right).
\]
The first term is the fraction of training examples whose margin is below a given value, and the second term is independent of the number of weak learners. Hence, in bounding the first term, one may ignore the complexity of the classifier.

It can be shown (Schapire and Freund, 2012) that the first term can be bounded as follows
\[
P_S[yf(x) \leq \theta] \leq e^\theta \sum_{t=1}^{T} Z_t,
\]
which, for small \(\theta\) tends to
\[
\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} \sqrt{1 - \gamma_t^2}.
\]
This well-known viewpoint provides additional justification for optimizing \(\prod_{t=1}^{T} Z_t\), as is done in the preceding section.

Experimental results

For our experiments, we first used datasets from the UCI repository, as shown in Table 1. The features and labels were collapsed into binary categories.

Experimental results, given in Figure 1, compare average generalization error rates over 20 trials, each with a random selection of training examples. Features are given costs uniformly at random on the interval \([0, 2]\). For comparison, AdaBoost was run for a number of rounds that gave lowest test error, irrespective of budget. This setup was chosen to compare directly against the results of Reyzin (2011) who also used random costs. We test on all the datasets Reyzin used, plus others.

Then, to study our algorithms on real-world data, we used the Yahoo! Webscope dataset 2, which includes feature costs (Xu, Weinberger, and Chapelle, 2012). The data set contains 519 features, whose costs we rescaled to costs to the set \{1, 5, 1, 2, 5, 10, 15, 20\}. Examples are query results labeled 0 (irrelevant) to 5 (perfectly relevant). We chose to collapse labels 1-5 to be binary label 1 (relevant) to test our algorithms. Results are given in Figure 2.

The most apparent conclusion from our experiments is that it is not only possible to improve upon AdaBoostRS by optimizing base learner selection during training, but that the improvement is dramatic. Further modifications of the basic AdaBoostBT tend to yield additional improvements.

AdaBoostBT\_Greedy often tends to perform better than the basic AdaBoostBT for small budgets, but it chooses base learners quite aggressively - a low cost base learner is extremely attractive at all rounds of boosting. This makes it possible that the algorithm falls into a trap, as in the sonar and ionosphere datasets, where we have a huge number of features (consequently, many features with cost close to zero). After 500 rounds of boosting, this approach still had not spent the budget of 2 because the same set of cheap features were re-used round after round leading to a deficient classifier. Similar behavior is seen for the ecoli and heart datasets.

AdaBoostBT\_Smoothed avoids this trap by considering the average cost instead. The appeal of cheap
Figure 1: Experimental results comparing our approaches to AdaBoostRS (Reyzin, 2011) and SpeedBoost (Grubb and Bagnell, 2012). Test error is calculated at budget increments of 2. The feature costs are uniformly distributed in the interval [0, 2]. The horizontal axis has the budget, and the vertical axis has the test error rate. AdaBoostRS test error rate uses the secondary vertical axis (on the right hand side) for all data sets except for heart. Error bars represent a 95% confidence interval.

base learners is dampened as the boosting round increases, with its limiting behavior to choose weak learners that maximize $\gamma$. Thus, we can see that using AdaBoostBT_Smoothed, while tending to perform worse than AdaBoostBT_Greedy for low budgets, tends to exceed its accuracy for larger budgets.
Table 1: Dataset sizes, numbers of features for training and test, and number of rounds when running the AdaBoost predictor.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>num features</th>
<th>training size</th>
<th>test size</th>
<th>AdaBoost rounds (optimized)</th>
<th>trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>ocr17</td>
<td>403</td>
<td>1000</td>
<td>5000</td>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>ocr49</td>
<td>403</td>
<td>1000</td>
<td>5000</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>splice</td>
<td>240</td>
<td>1000</td>
<td>2175</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td>census</td>
<td>131</td>
<td>1000</td>
<td>5000</td>
<td>880</td>
<td>100</td>
</tr>
<tr>
<td>breast cancer</td>
<td>82</td>
<td>500</td>
<td>199</td>
<td>500</td>
<td>400</td>
</tr>
<tr>
<td>ecoli</td>
<td>356</td>
<td>200</td>
<td>136</td>
<td>50</td>
<td>400</td>
</tr>
<tr>
<td>sonar</td>
<td>11196</td>
<td>100</td>
<td>108</td>
<td>99</td>
<td>400</td>
</tr>
<tr>
<td>heart</td>
<td>571</td>
<td>100</td>
<td>170</td>
<td>15</td>
<td>400</td>
</tr>
<tr>
<td>ionosphere</td>
<td>8114</td>
<td>300</td>
<td>51</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>webscope set2</td>
<td>519</td>
<td>7000</td>
<td>15439</td>
<td>500</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 2: Experimental results comparing our approaches to AdaBoostRS and SpeedBoost on the Yahoo! Webscope data set 2. Test error is calculated at budget increments of 2. Error bars represent a 95% confidence interval.

Comparing to Speedboost

We also compare to the classification algorithm SpeedBoost [Grubb and Bagnell, 2012] on all data sets, which was recently designed for a similar setting. SpeedBoost, works by choosing weak learners (together with a voting weight $\alpha$) so as to greedily optimize the improvement of a loss function per unit cost, until the budget runs out. To mirror Grubb and Bagnell (2012), as well as AdaBoost, we use exponential loss.

In our experiments, SpeedBoost performs almost identically to AdaBoostBT_Greedy. This phenomenon can be explained as follows. In AdaBoostBT_Greedy we find $\arg\min_{h \in \mathcal{H}} (1 - \gamma(h)^2)^{\frac{1}{\varepsilon(h)}}$, while in SpeedBoost, implicitly we find $\arg\min_{h \in \mathcal{H}} \frac{1 - \sqrt{1 - \gamma(h)^2}}{\varepsilon(h)}$. Since

$$\min_{h \in \mathcal{H}} (1 - \gamma(h)^2)^{\frac{1}{\varepsilon(h)}} = \max_{h \in \mathcal{H}} \frac{-\ln \sqrt{1 - \gamma(h)^2}}{\varepsilon(h)},$$

and the Taylor series of $-\ln(x)$ is

$$(1 - x) + \frac{1}{2} (1 - x)^2 - o((1 - x)^2),$$

we have when $\gamma(h)$ is close to 0 (the value toward which boosting drives edges by making hard distributions), the two functions SpeedBoost and AdaBoostBT_Greedy seek to optimize are very similar.

Moreover, both algorithms greedily optimize an objective function without considering the impact on future rounds. Hence, SpeedBoost falls into the same trap of copious cheap hypotheses as AdaBoostBT_Greedy. Note: the lines for SpeedBoost are dashed because they overlap with AdaBoostBT_Greedy.

Yet, our approach offers a number of benefits over SpeedBoost. Firstly, we have flexibility of explicitly considering future rounds, as AdaBoostBT_Smooth does, in many cases outperforming SpeedBoost—e.g. on the ecoli, heart, sonar, and ionosphere data. Second, computational issues (for a discussion, see [Grubb and Bagnell, 2012]) sur-
rounding choosing the best $\alpha$ is completely avoided in weak learner selection, with no need for a numerical approximation. Secondly, our approach yields more flexibility by limiting the problem to only weak learner selection. One can imagine using a hybrid approach, which selects between AdaBoostBT\_Greedy and AdaBoostBT\_Smooth depending on the dataset.

A note on decision trees

One straightforward method for making a budgeted predictor is to use decision trees, and we consider this approach in this section. Decision trees are a natural choice for budgeted predictor since after a tree is constructed, each test example only needs to go through one path of the tree in order the receive a label. Hence, it incurs a cost using features only on that particular path. One may even be tempted to think that simply using decision trees would be optimal for this problem. We experimentally show that this is not the case, especially for bigger budgets.

The problem with decision trees is that when one has more budget at hand and is able to grow larger trees, the trees begin to overfit (Figure 3). Hence this gap persists in our setting. Budgeted boosting algorithms continue to drive error rates down with higher budgets; decision trees do not.

Future Work

One direction for future work is to improve optimization for cost distributions with few cheap features. In addition, one could consider an adversarial cost model where cost and feature performance are strongly positively correlated, and analyze the extent to which optimization could help.

Other potentially promising approaches within our framework would be to boost weak learners other than decision stumps – for instance, boosting decision trees with the optimizations we suggest in Section 8 would likely outperform stumps, especially because decision trees show highest gains at small budgets (see Figure 3).

Finally, it would be worthwhile to study making other machine learning algorithms feature-efficient by incorporating budgets in their training.