Dissertation Research: It isn't enough to be smooth

Karen Glocer, Darrell Long, James Theiler

January 16, 2008

1 Executive Summary

The ISIS team in ISR-2 primarily uses supervised learning techniques to solve classification problems in imagery and therefore has a strong interest in finding linear classification algorithms that are both robust and efficient. Boosting algorithms take a principled approach to finding linear classifiers and they has been shown to be so effective in practice that they are widely used in a variety of domains.

In the real world, noisy data is ubiquitous, and this is equally true of the data in use at LANL. Unfortunately, it has been observed that the performance of the most common boosting algorithms deteriorates drastically with noise. The development of boosting algorithms that are robust to noise is only beginning. Virtually all boosting algorithms that have achieved some noise robustness do so by restricting the amount of weight that can be put on any example. This is called *smoothing* or *capping* in the literature.

In this proposal we present evidence that smoothing is not necessarily the optimal way to achieve noise robustness. We propose to research alternative mechanisms to smoothing to gain insight into the nature of noise robustness and to arrive at algorithms that achieve this efficiently.

2 Project Description

2.1 Background

Boosting methods have been used with great success in many applications like OCR, text classification, natural language processing, drug discovery, and computational biology [6]. For AdaBoost [4] it was frequently observed that the generalization error of the combined hypothesis kept decreasing after the training error had already reached zero [12]. This sparked a series of theoretical studies trying to understand the underlying principles that govern the behavior of ensemble methods [12, 1]. It became apparent that some of the power of ensemble methods lies in the fact that they tend to increase the margin of the training examples. This was consistent with the observation that AdaBoost works well on low-noise problems, such as digit recognition tasks, but not as well on tasks with high noise. On such tasks, a large margin on *all* training points cannot be achieved without adverse effects on the generalization error. This experimental observation was supported by the study of [12], where the generalization error of ensemble methods was bounded by the sum of two terms: the fraction of training points which have a margin smaller than some value ρ plus a complexity term that depends on the base hypothesis class and ρ. While this worst-case bound can only capture part of what is going on in practice, it nevertheless suggests that in some cases it pays to allow some points to have small margin or be misclassified if this leads to a larger overall margin on the remaining points.

To cope with this problem, it was necessary to construct variants of AdaBoost which trade off the fraction of examples with margin at least ρ with the size of the margin ρ. This was typically done by preventing the weighting maintained by the algorithm from concentrating too much on the most difficult examples. This idea is implemented in many algorithms including AdaBoost with soft margins [8], MadaBoost [3], ν-Arc [9, 7], SmoothBoost [13], LPBoost [2], and several others (see references in [6]). For some of these algorithms, significant improvements were shown compared to the original AdaBoost algorithm on high noise data.

In parallel there has been a significant interest in how the linear combination of hypotheses generated by AdaBoost is related to the maximum margin solution [1, 12, 2, 11, 10]. It was shown that AdaBoost generates a combined

hypothesis with a large margin, but not necessarily the maximum hard margin [8, 11]. This observation has led to the development of many new versions of AdaBoost that are provably able to maximize the margin [1, 5, 2, 10, 14, 11]. For AdaBoost* [10] and TotalBoost [14] it was shown that they converge in $2\ln(N/\delta^2)$ iterations to the maximum hard margin within precision δ, while the other algorithms had worse or no known convergence rates. However, such margin-maximizing algorithms are of limited interest for a practitioner working with noisy real-world data sets, as overfitting is even more problematic for such algorithms than for the original AdaBoost algorithm [1, 5].

In previous work, we combined these two lines of research into a single algorithm, called SoftBoost, that for the first time implements the soft margin idea in a practical boosting algorithm. SoftBoost finds a combined hypothesis with soft margin not smaller than the maximal soft margin minus δ in $O(\ln(N)/\delta^2)$ iterations.

From a theoretical point of view, the optimization problems underlying SoftBoost as well as LPBoost are appealing, since they directly maximize the margin of a (typically large) *subset* of the training data [9]. This quantity plays a crucial role in the generalization error bounds [12]:

$$P_{\mathcal{D}}[yf(x) \leq 0] \leq P_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{n}}\left(\frac{d\log^2(n/d)}{\theta^2} + \log(\frac{1}{\delta})\right)^{1/2}\right)$$

In the above equation, the left hand side is the generalization and the first term on the right corresponds closely to the soft margin. For clarity, n is the number of points in the training sample, θ represents the margin, and d is the VC-dimension. Informally, the VC-dimension is an indicator of the complexity of the hypothesis class. For a linear hypothesis class of dimension m, the VC-dimension is $m + 1$.

2.2 Maximizing the soft margin is not enough

Maximizing the margin has long been a proxy for good generalization, but it is only a proxy. Because one of the two terms in the above bound is so closely related to the soft margin, it has been commonly assumed that soft margin actually optimizes the above generalization bound, but closer investigation actually contradicts this argument. In a simple experiment, I demonstrate that an algorithm that does not maximize the soft margin actually does a better job of optimizing this bound.

I generated a synthetic data set by starting with a random matrix of 2000 rows and 100 columns, where each entry was chosen uniformly in $[0, 1]$. For the first 1000 rows, we added $1/2$ to the first 10 columns and rescaled such that the entries in those columns were again in $[0, 1]$. The rows of this matrix are our examples and the columns and their negation are the base hypotheses, giving us a total of 200 of them. The first 1000 examples were labeled $+1$ and the rest -1. This results in a well separable dataset. To illustrate how the algorithms deal with the inseparable case, we flipped the sign of a random 10% of the data set. We then chose a random 500 examples as our training set and the rest as our test set. In every boosting iteration we chose the base hypothesis which has the largest edge with respect to the current distribution on the examples. All parameters were chosen by cross-validation.

I then ran three smooth boosting algorithms on this synthetic data: LPBoost, SoftBoost, and SmoothBoost. LPBoost and SoftBoost maximize the soft margin directly, while SmoothBoost does so asymptotically. For comparison, I also ran BrownBoost, one of two known noisy boosting algorithms that does not fall into the catgory of smooth boosting. Figure 1 confirms that the smooth boosters maximize the soft margin while BrownBoost does not. [1] Figure 2 looks at the probability density function of the resulting margins of SoftBoost and BrownBoost. Also shows in the optimal soft margin for each algorithm. The first term of the bound is equivalent to the probability mass that lies below the soft margin.

The key insight results from plugging the results of SoftBoost and BrownBoost directly into the generalization bound. Because SoftBoost maximizes the soft margin while BrownBoost does not, we would expect the bound computed on the empirical data to be tighter for SoftBoost. Surprisingly, the bound is actually tighter for BrownBoost. For BrownBoost on this data, the generalization error is ≤ 21.5 while for SoftBoost, it is ≤ 22.2. More to the point, both algorithms have the same model complexity, but the fraction of points that lie below the soft margin is lower for BrownBoost. While the soft margin value depends on how far below the soft margin those points lie, the generalization bound does not. This is why SoftBoost has a higher soft margin but a looser generalization bound.

[1] The weights in the BrownBoost classifier to not sum to 1. To compare achieved margin fairly with the other algorithms, for the purposes of this plot the weights were normalized.

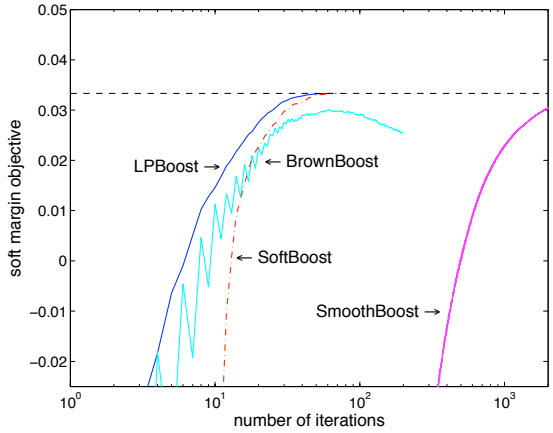

Figure 1: Soft margin objective vs. the number of iterations for LPBoost, SoftBoost, BrownBoost and SmoothBoost. BrownBoost, the only algorithm shown that does not use smoothing, is also the only one that does not maximize the margin.

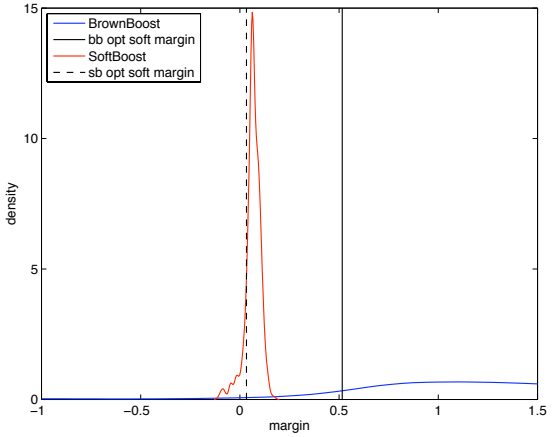

Figure 2: Margin probability density functions for SoftBoost and BrownBoost. Smooth boosting algorithms have most of their mass near the optimum soft margin.

2.3 Proposed Research

The above experiment suggests that although the vast majority of robust boosting algorithms rely on smoothing to make them robust to noise, there are other mechanisms that may be better, but they are poorly understood. Therefore, exploration of alternative mechanisms to make algorithms robust to noise seems like a promising avenue. BrownBoost and MartiBoost are the only two robust boosting algorithms that do not rely on smoothing. Some interesting directions suggested by the above experiments are:

- It appears that BrownBoost is robust to noise because it allows the algorithm to forget entirely about certain examples rather than to just limit their weight. There is some preliminary experimental evidence that *forgetting* is an effective mechanism, but it has not been studied.
- BrownBoost is the only forgetting algorithm currently in existence, and it does not maximize a margin or optimize any sort of objective function. It would not be surprising if an algorithm with the same robustness mechanism could be found that has additional good theoretical properties.
- Is it possible to use techniques similar to the ones used in SoftBoost to directly optimize the generalization bound?
- Is optimizing the generalization bound a sound approach in practice?

3 Personnel

Darrell Long	PI	Professor at UCSC
Karen Glocer	GSR	Graduate student at UCSC
James Theiler	LANL TSM	LANL Mentor

References

[1] L. Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11(7):1493–1518, 1999. Also Technical Report 504, Statistics Department, University of California Berkeley.

[2] A. Demiriz, K.P. Bennett, and J. Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46(1-3):225–254, 2002.

[3] C. Domingo and O. Watanabe. Madaboost: A modification of adaboost. In *Proc. COLT '00*, pages 180–189, 2000.

[4] Y. Freund and R.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, 1997.

[5] A.J. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *Proceedings of the Fifteenth National Conference on Artifical Intelligence*, 1998.

[6] R. Meir and G. Rätsch. An introduction to boosting and leveraging. In S. Mendelson and A. Smola, editors, *Proc. 1st Machine Learning Summer School, Canberra*, LNCS, pages 119–184. Springer, 2003.

[7] G. Rätsch. *Robust Boosting via Convex Optimization: Theory and Applications*. PhD thesis, University of Potsdam, Germany, December 2001.

[8] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, 2001.

[9] G. Rätsch, B. Schölkopf, A.J. Smola, S. Mika, T. Onoda, and K.-R. Müller. Robust ensemble learning. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 207–219. MIT Press, Cambridge, MA, 2000.

[10] G. Rätsch and M.K. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2131–2152, 2005.

[11] C. Rudin, I. Daubechies, and R.E. Schapire. The dynamics of adaboost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5:1557–1595, 2004.

[12] R.E. Schapire, Y. Freund, P.L. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[13] Rocco A. Servedio. Smooth boosting and learning with malicious noise. *Journal of Machine Learning Research*, 4:633–648, 2003.

[14] M.K. Warmuth, J. Liao, and G. Rätsch. Totally corrective boosting algorithms that maximize the margin. In *Proc. ICML '06*, pages 1001–1008. ACM Press, 2006.