JMLR: Workshop and Conference Proceedings vol (2012) 1–4

Does AdaBoost Always Cycle?

Cynthia Rudin RUDIN@MIT.EDU
Massachusetts Institute of Technology, MIT Sloan School of Management, Cambridge MA 02142 USA

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

Ingrid Daubechies INGRID@MATH.DUKE.EDU
Duke University, Department of Mathematics, 111 Physics, Durham, NC 27708 USA

Abstract

We pose the question of whether the distributions computed by AdaBoost always converge to a cycle.

The AdaBoost algorithm (Freund and Schapire, 1997) was designed to combine many "weak" hypotheses that perform slightly better than random guessing into a "strong" hypothesis that has very low error. Although extensively studied, some of AdaBoost's basic convergence properties are not fully understood. This open problem focuses on one of these, namely, the convergence of the distributions over training examples that are iteratively computed by the algorithm.

AdaBoost is shown in Fig. 1; see Schapire and Freund (2012) for further background. Briefly, we are given training examples $(x_1, y_1), \ldots, (x_m, y_m)$. On each of a sequence of rounds t, AdaBoost computes a distribution $\mathbf{D}_t$ over the training set which is used to select a weak hypothesis h_t from some space $\mathcal{H} = \{\hbar_1, \ldots, \hbar_N\}$, which we presume to be finite and closed under negation (so that $-h \in \mathcal{H}$ if $h \in \mathcal{H}$). To simplify the discussion, we assume that each weak hypothesis is selected *exhaustively*, meaning that h_t is chosen, among all $h \in \mathcal{H}$, to have minimum weighted error $\Pr_{i\sim\mathbf{D}_t}[h_t(x_i) \neq y_i]$, which is exactly equivalent to choosing h_t to have maximum weighted "correlation" r_t, as defined in the figure. The chosen weak hypotheses can eventually be combined into a final classifier H, as in the figure, although our focus here is only on the distributions $\mathbf{D}_t$.

Each distribution $\mathbf{D}_t$ can be viewed as a point in $\mathbb{R}^m$, or more specifically, on the probability simplex. AdaBoost, together with an exhaustive choice of weak hypotheses, can be regarded as defining a deterministic mapping from one distribution $\mathbf{D}_t$ to the next distribution $\mathbf{D}_{t+1}$.

Several authors (Rudin et al., 2004; Kutin, 2002; Amit and Blanchard, 2001) have independently noticed these distributions converging to cycles as t gets large. Such behavior is readily observed when the number of training examples and weak hypotheses is small. On the other hand, in the more realistic case of many examples and very many weak hypotheses, other authors have reported that AdaBoost's behavior can appear chaotic with respect to its distributions (Caprile et al., 2002).

In our experiments (Rudin et al., 2004), although the initial behavior may seem chaotic, the distributions tend to converge to a cycle. It is not known how to characterize the relationship of the examples and hypotheses to properties of the cycle, such as its length, which can vary substantially.

Boosting is often studied under a weak learning assumption, which, in our set-up, states that the correlations r_t are bounded away from zero (so that, for some $c > 0$, we have $r_t \geq c$ on every round t). When this assumption does *not* hold, it was shown by Collins et al. (2002) that the distributions

© 2012 C. Rudin, R.E. Schapire & I. Daubechies.

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$
set $\mathcal{H} = \{\hbar_1, \ldots, \hbar_N\}$ of weak hypotheses $\hbar_j : \mathcal{X} \to \{-1, +1\}$.
Initialize: $D_1(i) = 1/m$ for $i = 1, \ldots, m$.
For $t = 1, \ldots, T$:

- Train weak learner using distribution $\mathbf{D}_t$; that is, find weak hypothesis $h_t \in \mathcal{H}$ with maximum correlation $r_t \doteq \mathrm{E}_{i \sim \mathbf{D}_t}[y_i h_t(x_i)]$.
- Choose $\alpha_t = \frac{1}{2} \ln((1 + r_t)/(1 - r_t))$.
- Update, for $i = 1, \ldots, m$: $D_{t+1}(i) = D_t(i) \exp(-\alpha_t y_i h_t(x_i))/Z_t$
  where Z_t is a normalization factor (chosen so that $\mathbf{D}_{t+1}$ will be a distribution).

Output the final hypothesis: $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

Figure 1: The boosting algorithm AdaBoost.

$\mathbf{D}_t$ must converge to a single, uniquely-defined point (a degenerate cycle of length one). When the weak learning assumption *does* hold, the distributions cannot converge to a single point, but they still may converge to a cycle. Thus, the open problem is concerned with this latter case only.

If it were possible to show that AdaBoost's distributions always converge to a cycle, and if one could actually find the cycle (either analytically or numerically), we might be able to substantially speed up the algorithm by "jumping" to its asymptotic behavior. Or we might be able to directly solve for AdaBoost's asymptotic "minimum margin," perhaps yielding direct insight into its ability to generalize to unseen training examples.

The mapping of $\mathbf{D}_t$ to $\mathbf{D}_{t+1}$ induced by AdaBoost can be greatly simplified. To do so, we define the $m \times N$ matrix $\mathbf{M}$ by $M_{ij} = y_i \hbar_j(x_i)$, thus encoding which weak hypotheses $\hbar_j \in \mathcal{H}$ are correct on which training examples (x_i, y_i). More abstractly, $\mathbf{M}$ can be viewed as an arbitrary $\{-1, +1\}$-valued matrix. Given $\mathbf{D}_t$, AdaBoost's computation of $\mathbf{D}_{t+1}$ can then be written equivalently as:

1. $j_t = \text{argmax}_j(\mathbf{D}_t^\top \mathbf{M})_j$.
2. $r_t = (\mathbf{D}_t^\top \mathbf{M})_{j_t}$.
3. $D_{t+1}(i) = D_t(i)/(1 + r_t\, M_{ij_t})$ for $i = 1, \ldots, m$.

In step 1, a column j_t is selected with maximum correlation, corresponding to the choice of weak hypothesis $h_t = \hbar_{j_t}$. In step 2, the correlation r_t is computed. And in step 3, the new distribution $\mathbf{D}_{t+1}$ is computed, here written in an explicit form that does not require further normalization.

We say that the distributions $\mathbf{D}_t$ converge to a cycle if there exist "cycle points" (distributions) $\hat{\mathbf{D}}_1, \ldots, \hat{\mathbf{D}}_\ell$ such that $\mathbf{D}_{k\ell+b} \to \hat{\mathbf{D}}_b$ as integer $k \to \infty$, for $b = 1, \ldots, \ell$. Thus, the open problem is to determine if, for every matrix $\mathbf{M}$, the distributions $\mathbf{D}_t$ necessarily converge to a cycle.

Note that the maximizing column in step 1 may not be unique, in which case it is necessary to assume that ties are broken in some consistent fashion; for concreteness, let us suppose they are broken by selecting the column whose index j is smallest. Further, this step breaks the probability simplex of distributions into regions based on which hypotheses would be selected for which distributions; the ties occur at the boundary of such regions. As a result of this step, the iterative map is highly discontinuous. This lack of continuity is the cause of much of the difficulty in working with this problem since it means that many of the classical results on dynamical systems are inapplicable. For instance, this is the primary reason why the well-known "Period Three Implies Chaos" result (Li and Yorke, 1975) does not apply. (In fact, there are matrices $\mathbf{M}$ where every distribution must converge to a 3-cycle.)

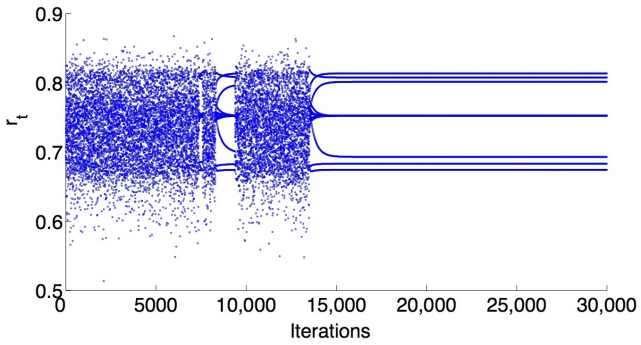

Figure 2: A plot of r_t over 30,000 iterations of AdaBoost on a small matrix $\mathbf{M}$.

From our previous studies (Rudin et al., 2004), we know there are some simple matrices $\mathbf{M}$ for which the distributions must converge to a cycle, and the iterated map provably forms a contraction in which nearby distributions $\mathbf{D}_t$ must get closer to the cycle points over time. (There are multiple possible cycles, but every possible distribution must converge to one of them.) Also, there is sometimes an analytical expression for these cycle points, and sometimes it is possible to prove there exists a unique solution for the cycle points if there is no closed-form solution.

In experiments on small matrices $\mathbf{M}$, we have observed cycles of many different lengths, including odd and even lengths. Sometimes, AdaBoost takes a very long time to converge to a cycle. If one of the cycle points is close to the boundary between regions of the simplex, as the distribution is converging to the cycle, it could cross the boundary. At that point the distributions could map to a different part of the simplex altogether, and leave the region of attraction. This is illustrated in Fig. 2 where one of AdaBoost's parameters (r_t) is plotted over 30,000 iterations of AdaBoost. The apparent lines in the figure are made as AdaBoost alternates between a small number of possible values of r_t as it cycles. Around iteration 9,000, the weight vector crosses one of the regions in the simplex and no longer follows its previous cycle. Eventually, it finds this cycle and converges again.

The open problem is to prove or disprove that AdaBoost's distributions $\mathbf{D}_t$ converge to a cycle in all cases, that is, for every $\{-1, +1\}$-valued matrix $\mathbf{M}$. A reward of \$100 is offered for a complete and general resolution of this problem.

References

Yali Amit and Gilles Blanchard. Multiple randomized classifiers (MRCL). Unpublished manuscript, 2001.

Bruno Caprile, Cesare Furlanello, and Stefano Merler. Highlighting hard patterns via AdaBoost weights evolution. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, pages 72–80. Springer, 2002.

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

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.

Samuel Kutin. *Algorithmic stability and ensemble-based learning*. PhD thesis, University of Chicago, 2002. pages 249–250.

Tien-Yien Li and James A Yorke. Period three implies chaos. *The American Mathematical Monthly*, 82(10): 985–992, December 1975.

Cynthia Rudin, Ingrid Daubechies, and Robert E. Schapire. The dynamics of AdaBoost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5:1557–1595, Dec 2004.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.