F_046

Boosting Using Classifiers with Nearly One-Sided Error

Kohei Hatano†

1 Introduction

Boosting is a powerful tool in the machine learning literature and it has been extensively studied over a decade. For binary classification problems, a typical boosting algorithm works in an iterative fashion: For each iteration $t = 1, \ldots, T$, a boosting algorithm constructs a distribution D_t over the data, and learns a base classifier, or a *weak hypothesis* h_t whose error is slightly less than $1/2$ w.r.t. D_t. The final classifier is the weighted linear combination of weak hypotheses $h_1, \ldots, h_T$, which is supposed to be accurate enough.

Assuming that each weak hypothesis h_t have error less than $1/2 - \gamma$ $(0 < \gamma < 1/2)$, Boost-by-Majority algorithm [2] and other successors (e.g., AdaBoost [3]) can construct a combined hypothesis with error less than ε using $O((1/\gamma^2)\log(1/\varepsilon))$ weak hypotheses. In [2], it is also shown that $\Omega(1/\gamma^2 \log(1/\varepsilon))$ weak hypotheses are necessary when a combined hypothesis is represented with a majority vote of weak hypotheses.

On the other hand, a $(1-\varepsilon)$-accurate classifier can be constructed by $O((1/\gamma)\log(1/\varepsilon))$ weak hypotheses when they have *one-sided error*, i.e., their positive predictions are always correct (Of course, one can also consider the negative version of one-sided error) [6, 4].

Although the improvement by the factor $1/\gamma$ is nontrivial, hypotheses with one-sided error are rarely available in practice. However, it is more likely that hypotheses have "nearly" one-sided error, in other words, they might have low false-positive error. For example, suppose that we'd like to predict whether the subject of an article in the newspaper is economy or not by using hypotheses associated with words. Then a hypothesis that predicts positive if an article contains the word "stock" would have low false positive error.

In this paper, we investigate a boosting scheme that can take advantage of the situation where weak hypotheses have nearly one-sided error. We use InfoBoost [1] (a simple version of [7]) and show that, under the assumption that each weak hypothesis have error less than $1/2 - \gamma$ and its false positive error is at most τ times its error $(0 \leq \tau \leq 1/2)$, one can construct a $(1-\varepsilon)$-accurate hypothesis using $O\left((\tau/\gamma + (1-2\tau)/\gamma^2)\log(1/\varepsilon)\right)$ weak hypotheses. This bound interpolates previous ones for the cases of one-sided and two-sided error. We also show an application of InfoBoost for learning a class of linear threshold functions with large constant biases, which includes classes of boolean functions such as disjunctions and r-of-k functions [5].

†Department of Informatics, Kyushu University

2 Preliminaries

Let X be the *instance space* and $Y = \{-1, +1\}$ be the set of labels. A pair $(\boldsymbol{x}, y) \in X \times Y$ is called an *example*. The learner is given a multiset S of m examples, $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$, where each example $(\boldsymbol{x}_i, y_i)$ is drawn independently from an unknown distribution D over $X \times Y$. A function $h : X \to Y$ is called a *hypothesis* and let H' be a set of hypotheses. Given S and H', the learner' goal is to output a hypothesis $h \in H'$ whose *error* $\mathrm{err}_D(h) \stackrel{\mathrm{def}}{=} \Pr_D\{h(\boldsymbol{x}) \neq y\}$ is small as possible.

In particular, we assume that the learner is given a set H of base hypotheses and outputs a hypothesis from the set $\mathrm{conv}_T(H)$ of linear combinations of T hypotheses in H for some $T \geq 1$. Based on VC-theory, the following statement holds (See e.g., [3]): With high probability, it holds for any $h \in \mathrm{conv}_T(H)$ that

$$\mathrm{err}_D(h) \leq \widehat{\mathrm{err}}_S(h) + \tilde{O}(\sqrt{Td_H/m}),$$

where $\widehat{\mathrm{err}}_S(h) \stackrel{\mathrm{def}}{=} |(\boldsymbol{x}_i, y_i) \in S \mid h(\boldsymbol{x}_i) \neq y_i\}|/m$, and d_H is the *VC dimension* of H. So, given H of fixed VC-dimension, a strategy to get a hypothesis with small error is to find a hypothesis whose size and the training error is small. For any distribution D over $X \times Y$ and any hypothesis h, let

$$\mathrm{tp}_D(h) \stackrel{\mathrm{def}}{=} \Pr_D\{h(\boldsymbol{x}) = +1, y = +1\}$$

$$\mathrm{fn}_D(h) \stackrel{\mathrm{def}}{=} \Pr_D\{h(\boldsymbol{x}) = -1, y = +1\},$$

$$\mathrm{fp}_D(h) \stackrel{\mathrm{def}}{=} \Pr_D\{h(\boldsymbol{x}) = +1, y = -1\}, \text{ and}$$

$$\mathrm{tn}_D(h) \stackrel{\mathrm{def}}{=} \Pr_D\{h(\boldsymbol{x}) = -1, y = -1\}.$$

Note that $\mathrm{err}_D(h) = \mathrm{fp}_D(h) + \mathrm{fn}_D(h)$. Now we give a formal definition of a weak hypothesis with low false-positive error:

Definition 1. A hypothesis h is a (τ, γ)-*weak hypothesis* w.r.t. D if $\mathrm{fp}_D(h) = \tau(\mathrm{err}_D(h))$ and $\mathrm{err}_D(h) = 1/2 - \gamma$.

3 Our Analysis

We apply InfoBoost for the case where (τ, γ)-weak hypotheses are available. A description of InfoBoost is given in Figure 1. The training error of h_{final} output by InfoBoost is bounded as follows:

InfoBoost

Given: $S = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m)\}$;

begin

1. $D_1(i) = 1/m$ $(i = 1, ..., m)$;
2. **For** $t = 1, \ldots, T$ **do**

   a) Learn $h_t \in H$ w.r.t. D_t;

   b) $\alpha_{t,+} = \frac{1}{2} \ln \frac{\mathrm{tp}_{D_t}(h_t)}{\mathrm{fp}_{D_t}(h_t)}$; $\alpha_{t,-} = \frac{1}{2} \ln \frac{\mathrm{tn}_{D_t}(h_t)}{\mathrm{fn}_{D_t}(h_t)}$;

   c) Let $\alpha_t(x) = \begin{cases} \alpha_{t,+}, & x \geq 0, \\ \alpha_{t,-}, & x < 0. \end{cases}$

   d) Update the distribution as follows:

$$D_{t+1}(i) = \frac{D_t(i) e^{-y_i \alpha_t(h_t(\boldsymbol{x}_i)) h_t(\boldsymbol{x}_i)}}{Z_t},$$

   where Z_t is the constant s.t. $\sum_{i=1}^{m} D_{t+1}(i) = 1$;

   end-for

3. Output $h_{\mathrm{final}}(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t(h_t(\boldsymbol{x}_i)) h_t(\boldsymbol{x}_i)$;

end.

Figure 1: InfoBoost

Theorem 1 ([7, 1]).

$$\widehat{\mathrm{err}}_S(h_{\mathrm{final}}) \leq \prod_{t=1}^{T} Z_t,$$

where $Z_t = p_t \sqrt{1 - \gamma_{t,+}^2} + (1 - p_t) \sqrt{1 - \gamma_{t,-}^2}$, $p_t = \Pr_{D_t}[h_t(\boldsymbol{x}_i) = +1]$ and $\gamma_{t,\pm} = \mathbf{E}_{D_t}[y_i h_t(\boldsymbol{x}_i) | h_t(\boldsymbol{x}_i) = \pm 1]$.

By concavity of the function $g(x) = \sqrt{1 - x}$, we obtain the following lemma.

Lemma 1. For any $t \geq 1$, $Z_t \leq e^{\frac{p_t \gamma_{t,+}^2 + (1 - p_t) \gamma_{t,-}^2}{2}}$.

By using Lemma 1 and the fact that $\Pr_{D_t}\{y_i = +1\} = 1/2$ for $t \geq 2$, we prove the following theorem:

Theorem 2. Given that, for each $t = 2, \ldots, T$, h_t is a (τ_t, γ_t)-weak hypothesis w.r.t. D_t and $\gamma_t \geq \gamma$ and $\tau_t \leq \tau$, InfoBoost outputs h_{final} with $\widehat{\mathrm{err}}_S(h_{\mathrm{final}}) \leq \varepsilon$ by using $T = O((\tau/\gamma^2 + (1 - 2\tau)/\gamma) \log(1/\varepsilon))$ weak hypotheses.

We summarize our analysis in Table 1.

4 Learning Linear Threshold Functions With Biases

In this section, we show that InfoBoost can learn a class of linear threshold functions with biases efficiently. For any ρ and σ $(0 < \sigma, \rho < 1)$, let $LTF_N(\rho, \sigma)$ be the set of the following linear threshold functions over $\{-1, +1\}^N$:

$$f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{\alpha} \cdot \boldsymbol{x} + 1 - \delta),$$

where $||\boldsymbol{\alpha}||_1 = 1$, and $|\boldsymbol{\alpha} \cdot \boldsymbol{x} + 1 - \delta| \geq \rho$ for each $\boldsymbol{x} \in \{-1, +1\}^N$.

$\tau = 0$ (one-sided)	$0 < \tau < 1/2$ (nearly one-sided)	$\tau = 1/2$ (two-sided)
$O\left(\frac{1}{\gamma}\right)$	$O\left(\frac{\tau}{\gamma^2} + \frac{1-2\tau}{\gamma}\right)$	$O\left(\frac{1}{\gamma^2}\right)$

Table 1: Number of (τ, γ)-weak hypotheses sufficient for boosting (The factor $\log(1/\varepsilon)$ is omitted).

Let H be the set of N boolean literals over $\{-1, +1\}^N$. Then we prove the existence of (τ, γ)-weak hypotheses for $LTF_N(\rho, \sigma)$.

Lemma 2. For any $f \in LTF_N(\rho, \delta)$ and any distribution D over $\{-1, +1\}^N$ for which $\Pr_D\{f(\boldsymbol{x}) = +1\} = 1/2$, there exists a (τ, γ)-weak hypothesis $h \in H$ w.r.t. D and f such that (i) $\gamma \geq \rho/(2\sqrt{\sigma})$, or (ii) $\gamma \geq \rho/2$ and $\tau \leq 2(\delta - \rho)$.

By applying Theorem 2 and Lemma 2, we prove the following theorem.

Theorem 3. Fix any $f \in LTF_N(\rho, \sigma)$. Given a multiset $S = \{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}$ of m examples, InfoBoost outputs h_{final} with $\widehat{\mathrm{err}}_S(h_{\mathrm{final}}) \leq \varepsilon$ by using $T = O((\delta/\rho^2) \log(1/\varepsilon))$ weak hypotheses.

Say, for $\delta = 2\rho$, $T = O((1/\rho) \log(1/\varepsilon))$ are weak hypotheses are sufficient. Our bound is better than the one $O((1/\rho^2) \log(1/\varepsilon))$ obtained by typical boosting algorithms such as AdaBoost.

References

[1] J. A. Aslam. Improving algorithms for boosting. In *Proc. 13th Annu. Conference on Comput. Learning Theory*, pages 200–207, 2000.

[2] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.

[3] 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.

[4] K. Hatano and M. K. Warmuth. Boosting versus covering. In *Advances in Neural Information Processing Systems 16*, 2003.

[5] K. Hatano and O. Watanabe. Learning r-of-k functions by boosting. In *Proceedings of the 15th International Conference on Algorithmic Learning Theory*, pages 114–126, 2004.

[6] B. K. Natarajan. *Machine Learning: A Theoretical Approach*. Morgan Kaufmann, San Mateo, CA, 1991.

[7] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.