Margin Theory for Voting Classifiers

王立威

北京大学智能科学系

Outline

- A Brief Review of AdaBoost and Margin Theory.
- Breiman's Doubt on the Margin Explanation.
- Related Work and Improvements.
- Our Results: Equilibrium Margin, Sharper Margin Bounds.

A Brief Review of AdaBoost and Margin Theory

Brief Review of AdaBoost

- Adaboost produces a linear combination (also called voting) of a number of *base* classifiers $h_i(x)$.

$$f(x) = \sum_i \alpha_i h_i(x) \qquad \alpha_i \geq 0, \ \sum_i \alpha_i = 1.$$

- Adaboost has demonstrated excellent experimental performance both on benchmark datasets and real applications.

- Mystery of adaboost: the test error of the combined classifier usually continuously decreases as its size becomes very large and even after the training error is zero. Not over-fitting? Contradict to Occam's razor?

- Schapire et al (Ann. Stat. 1998) developed a **margin theory** to explain the empirical observation of adaboost.

- We consider only binary classification problems. An example is $(x, y), \quad y \in \{-1,1\}$.

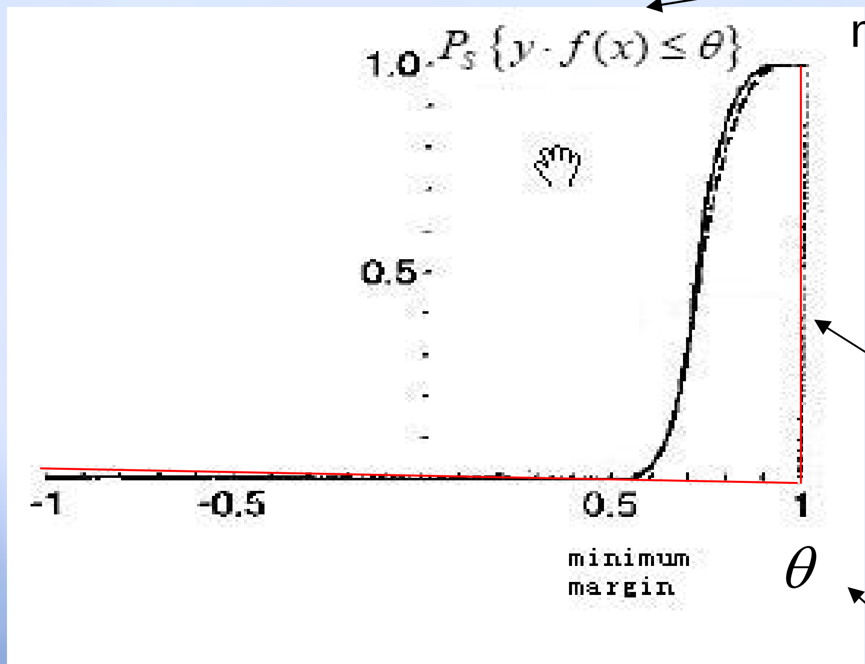
- Each base classifier only output -1 or 1, so the range of the output of the voting classifier is [-1,1]:

$$h_i(x) \in \{-1,\ 1\}. \qquad f(x) \in [-1,\ 1].$$

- If $y \cdot f(x) > 0$, the classification is correct, and makes an error otherwise.

- The margin of an example (x, y) is $y \cdot f(x)$ which represents the confidence of this classification result.

- The adaboost algorithm has the ability to make most of the training examples to have large margins.
- The distribution of the margins of all training examples are called margin distribution.



- Adaboost tends to make the margin distribution "good".

Margin Theory

- Schapire et al's margin theory:
  - The (upper bound of) generalization error of a voting classifier depends on the **margin distribution,** when the number of training examples and the number (or VC dimension) of the base classifier are fixed.

generalization error

\# training examples

Size of the set of base classifiers

with prob. at least $1-\delta$

$$\forall_\delta \quad P_D\{y\cdot f(x)\le 0\}\le \inf_{\theta\in(0,1]}\left\{P_S\{y\cdot f(x)\le\theta\}+O\left(\frac{1}{\sqrt{n}}\left(\frac{\log n\cdot\log|H|}{\theta^2}+\log 1/\delta\right)^{1/2}\right)\right\}.$$

  - If most of the training examples have large margins, then the generalization error has a small upper bound.

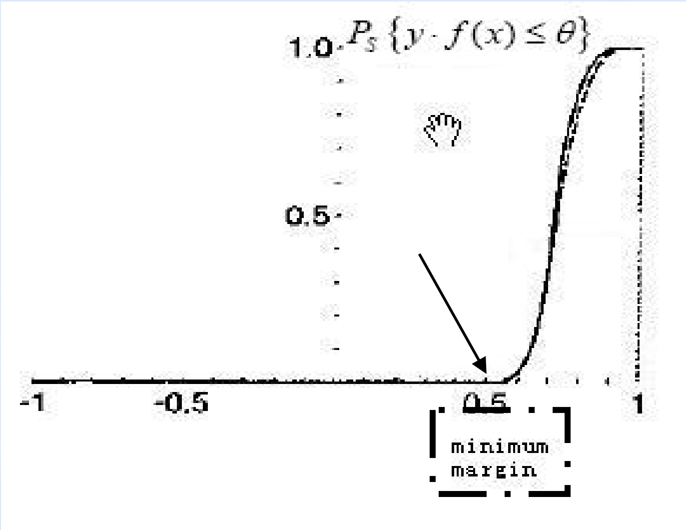
Breiman’s Doubt on the Margin Explanation

Crisis: Breiman's Doubt

- Breiman's margin bound:
  - Using an improved argument of Schapire's bound, Breiman gave a sharper upper bound of the generalization error of voting classifiers.
  - It says: the (upper bound of) generalization error of a voting classifier depends on the **minimum margin,** when the number of training examples and the number (or VC dimension) of the base classifier are fixed.

- Minimum margin is the maximum value of the margins at which the training error is zero.

$$\theta_0 = \sup\left\{\theta \in (0,1],\ \ P_S\left\{y \cdot f(x) \le \theta\right\} = 0\right\}.$$



- The bound: $\forall_\delta \quad P_D\left\{y \cdot f(x) \le 0\right\} \le O\left(\frac{1}{n}\left(\frac{\log n \cdot \log|H|}{\theta_0^2} + \log 1/\delta\right)\right).$

- Breiman's vs. Schapire's bound:

$$O(\frac{\log n}{n}) \quad \text{vs.} \quad O(\sqrt{\frac{\log n}{n}}).$$

  - Breiman's is better.
  - Seems that minimum margin governs the generalization error.

- Arc-gv algorithm:
  - Arc-gv is also boosting type algorithm, but the voting classifier it generates provably maximizes the minimum margin.
  - According to Breiman's margin bound, arc-gv should have better performance than adaboost.

- Surprise: arc-gv is consistently worse than adaboost in all the experiments!

- Breiman's doubt:
  - Margin theory does not explain why adaboost works so well, margin has noting to do with the generalization error. (Perhaps because the theory only gives upper bounds?)

- **Margin theory is in danger!**

Related Work and Improvements

Reyzin and Schapire's Analysis on Arc-gv (ICML06)

- A closer look at the margin bounds:

Size of the set of base classifiers

$$\forall_\delta \quad P_D\{y \cdot f(x) \le 0\} \le$$

$$\inf_{\theta \in (0,1]} \left\{ P_S\{y \cdot f(x) \le \theta\} + O\left(\frac{1}{\sqrt{n}} \left(\frac{\log n \cdot \log |H|}{\theta^2} + \log 1/\delta \right)^{1/2} \right) \right\}.$$

  - Generalization error also depends on the complexity of the base classifiers.

- Are the base classifiers in arc-gv and adaboost have the same complexity?

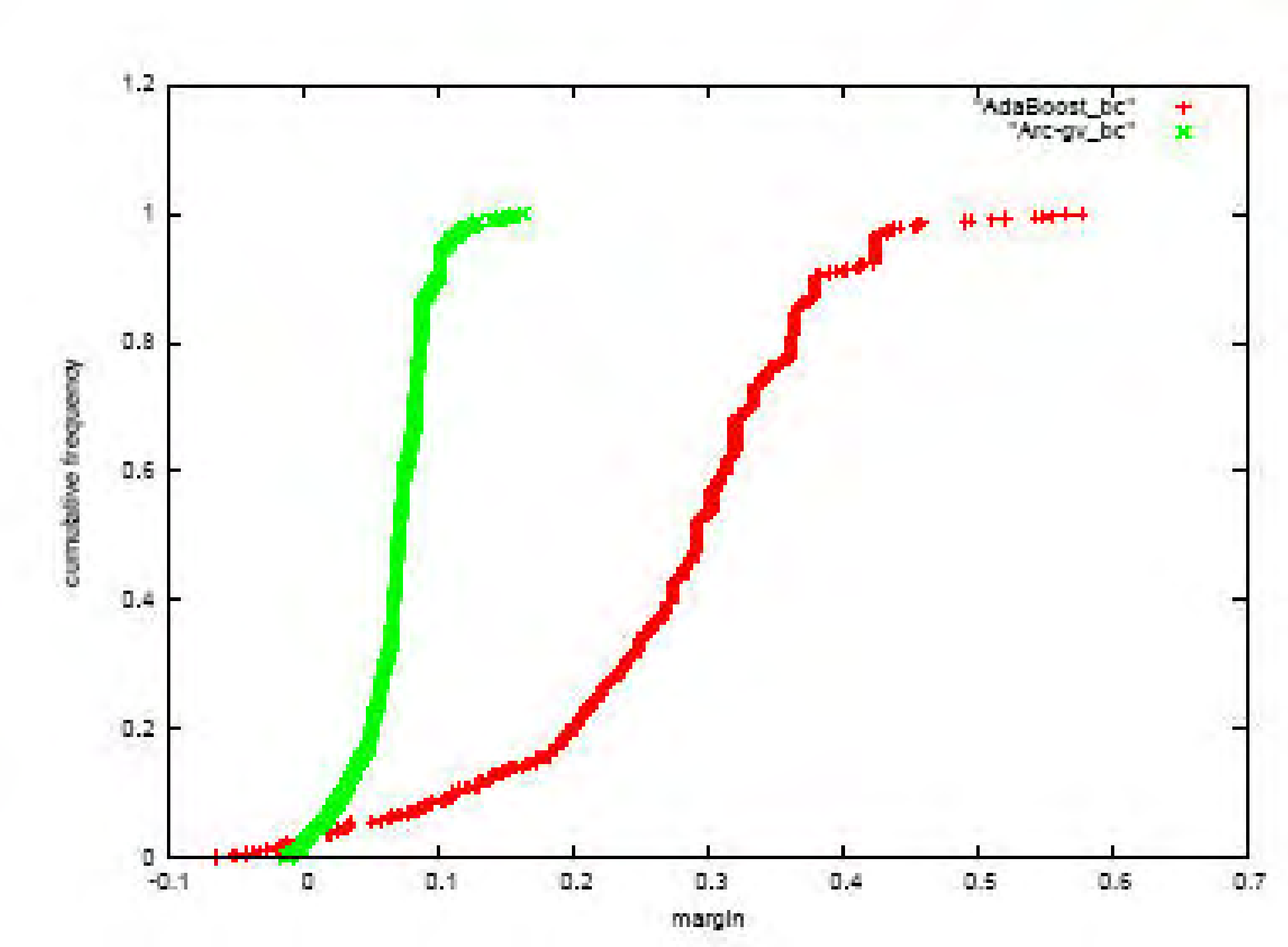
- How Breiman controlled the complexity of the base classifiers in his experiments for comparing arc-gv and adaboost?
  - He used decision trees (CART) as the base classifiers;
  - He controlled the complexity of the base classifiers by always choosing decision trees of a fixed size (number of nodes of the tree).

- But, Reyzin and Schapire found that trees produces by arg-gv are significantly deeper than those produces by adaboost!

- Not only size, but depth are complexity measures of trees.

- Breiman’s experiment was unfair! Arc-gv used more complex base classifiers.

- Controlling Classifier Complexity:
  - Using decision stumps (simplest decision, depth=1) as the base classifier. They all have the same complexity for any measure.
  - Observation:
    - Arc-gv is still worse than adaboost.
    - Although arc-gv produces larger minimum margin, the margin distribution is not as “**good**” as that adaboost generates.



(Reyzin & Schapire ICML06)

Improved Margin Bounds

- Koltchinskii and Panchanko's bound (Ann. Stat. 2005):
  - There exists an absolute constant K, such that for all $\varepsilon > 0$

$$\forall_{\delta} \quad P_D\left\{y \cdot f(x) \le 0\right\} \le \inf_{\theta \in (0,1]} \left\{ \begin{array}{l} (1+\varepsilon) P_S\left\{y \cdot f(x) \le \theta\right\} + \\ (2+\varepsilon+\frac{1}{\varepsilon}) \cdot K \cdot \left(\frac{1}{n} \left(\frac{\log n \cdot \log |H|}{\theta^2} + \log {}^1\!/\!_\delta \right) \right) \end{array} \right\}.$$

  - Comments on this bound:
    - O(log(n)/n) bound, the same as Breiman's, but with unknown constants. Can not compared to other bound in finite example situation.

Some Thoughts

- If we can propose margin bound that is provebly better than Breiman's, and the generalization error depends on the "whole" margin distribution, but not the minimum margin, then we can answer to Breiman's doubt!
- This motivates of our work!

Equilibrium Margin: Sharper Margin Bounds

Our Results:

- Sharper Margin Bound:

$$\forall_{\delta} \quad P_D\left\{y \cdot f(x) \le 0\right\} \le O\left(\frac{1}{n}\left(\frac{\log n \cdot \log|H|}{\hat{\theta}^2} + \log {}^1\!/\!_{\delta}\right)\right),$$

Equilibrium Margin:

$$\hat{\theta} = \sup\left\{\theta \in (0,1],\ \ P_S\left\{y \cdot f(x) \le \theta\right\} < \frac{4}{n}\left(\frac{16}{\theta^2}\log n \log|H| + \log {}^1\!/\!_{\delta}\right)\right\}.$$

- Note that equilibrium margin is always larger than minimum margin!

- Comparison to Breiman's minimum margin bound:
  - Breiman's:

$$\forall_{\delta} \quad P_D\left\{y \cdot f(x) \leq 0\right\} \leq O\left(\frac{1}{n}\left(\frac{\log n \cdot \log\left|H\right|}{\theta_0^2} + \log {}^{1}\!/_{\delta}\right)\right).$$

  - Ours:

$$\forall_{\delta} \quad P_D\left\{y \cdot f(x) \leq 0\right\} \leq O\left(\frac{1}{n}\left(\frac{\log n \cdot \log\left|H\right|}{\hat{\theta}^2} + \log {}^{1}\!/_{\delta}\right)\right),$$

  - Our bound based on equilibrium margin is better than Breiman's (with some loss in constants).

- Further improvement:
  - Using inverse function of Bernoulli relative entropy:

$$\forall_\delta \quad P_D\left\{y \cdot f(x) \le 0\right\} \le \frac{n+1}{n^2}\log^2|H| +$$

$$\inf_{t \ge 0}\left\{ D^{-1}\left(t, \frac{1}{n}\left(\begin{array}{l} 6\log\frac{4}{\hat{\theta}(t)} + 3\log\log\left(\frac{n}{\log|H|}\right) + \\ \frac{16}{\hat{\theta}(t)^2}\log\left(\frac{n}{\log|H|}\right)\log|H| + \log{}^{n}\!/_{\delta} \end{array} \right)\right)\right\},$$

$$\hat{\theta}(t) = \sup\left\{\theta \in (0,1],\ \ P_S\left\{y \cdot f(x) \le \theta\right\} < t\right\}.$$

  - It can be shown that this bound is consistently better than Breiman's up to a log(n)/n term, which can be ignored.

Summarize

- We give sharper margin bound for adaboost (and all voting classifiers), in which the generalization error is characterized by a new margin measure called equilibrium margin.

- The equilibrium margin bound is consistently better than Breiman's minimum margin bound (only up to a log(n)/n term).

What Does Our Bound Imply?

- Minimum margin is not the characteristic of the generalization error for adaboost.
- The fact that arc-gv produces larger minimum margin yet worse performance than adaboost can be explained by our equilibrium margin bound, because adaboost generates “better” margin distribution and larger equilibrium margin.

Thanks