Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods

By Schapire, Freund, Bartlett and Lee

Presented by: Sweta Sinha

Overview

- Introduction
- Background
- Bagging
- Idea of boosting
- Error Analysis of boosting
- Generalization error analysis based on margin
- Relation to Bias – variance theory
- Experiments
- Conclusions

Introduction

- Learning algorithm operates on given set of instances to produce a classifier
- Goal is to find classifier with low generalization error
- Focus on algorithm which achieve high accuracy by voting
  - Base classifier – each classifier combined in vote
  - Combined classifier – final vote classifier
- Boosting and bagging two common method
- Analysis of prediction error

Background

- Valiant'84
  - introduced theoretical PAC model for studying machine learning
- Kearns&Valiant'88
  - Can weak learner be "boosted" into accurate algorithm?
- Schapire'89 , Freund'90
  - first polynomial-time boosting algorithms
- Freund&Schapire '95
  - introduced AdaBoost algorithm
  - strong practical advantages over previous boosting algorithms
- continuing development of theory & algorithms:

  Schapire,Freund,Bartlett&Lee '97 Schapire&Singer '98
  Breiman '97 Mason, Bartlett&Baxter '98
  Grive and Schuurmans'98 Friedman, Hastie&Tibshirani '98

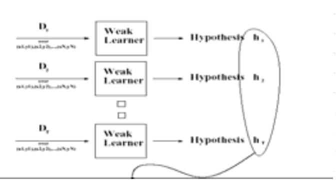
Bagging

- Combined the prediction of several classifiers
- Repeatedly
  - Samples data with replacement from the training set
  - Train a new classifier on the sample data
- The predictions of the classifier are combined by majority vote

Bagging works by reducing the <u>variance</u> part

Boosting

- Popular method of producing ensemble
- General method of converting rule of thumbs into highly accurate prediction rule.
- "Weak" learning algorithm combines to consistently find hypothesis with lower error



Final Hypothesis: f(h1, h2, h3, …ht)

Idea of Boosting

- Examine the training set X = {(x1,y1), ..(xm,ym)}
  $y_i \in \{-1,+1\}$ correct label of instance $x_i \in X$
- Derive some rough rule of thumb
- Reweight the sample – concentrate on "hard" cases for the previous rule
- Derive a second rule of thumb
- Repeat T times …
- Combine the rules of thumb into a single accurate rule

Boosting works by reducing the <u>bias</u> part

Boosting: Reweighing the sample

- for $t = 1,\ldots,T$:
  - construct distribution D_t on $\{1,\ldots,m\}$
  - Find <u>weak hypothesis</u> ("rule of thumb")

    $$h_t : X \to \{-1,+1\}$$

    with small error ε_t on D_t:

    $$\varepsilon_t = \Pr_{D_t}[h(x_i) \neq y_i] = \sum_{i: h_t(x_i) \neq y_i} D_t(i)$$
- output <u>final hypothesis</u> H_{final}

Ada Boost

- constructing D_t:
  - $$D_1(i) = \frac{1}{m}$$
  - given D_t and h_t:

$$D_{t+1} = \frac{D_t}{Z_t} \cdot \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$

$$= \frac{D_t}{Z_t} \cdot \exp(-\alpha_t \cdot y_i \cdot h_t(x_i))$$

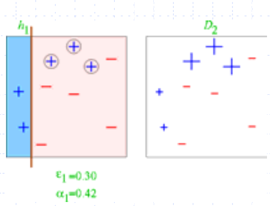
where: Z_t = normalization constant

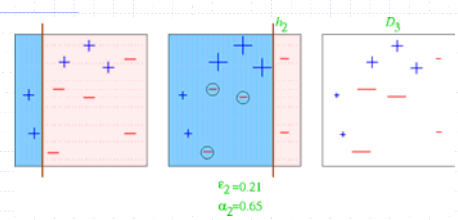
$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) > 0$$

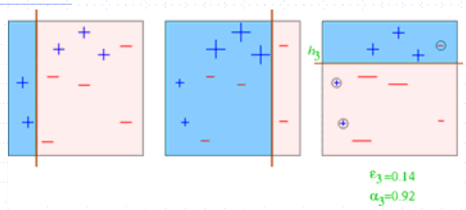
- final hypothesis: $$H_{\text{final}}(x) = \operatorname{sgn}\left(\sum_t \alpha_t h_t(x)\right)$$

Illustrative Example

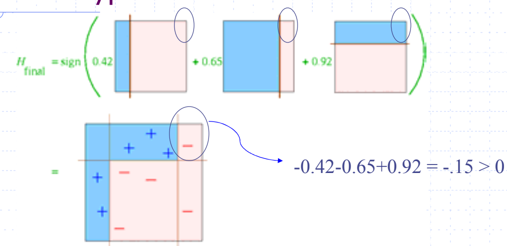
Example cont…







Final Hypothesis



Overview

- Introduction
- Idea of Boosting
- Error Analysis of boosting
- Generalization error analysis based on margin
- Experiments
- Conclusions

Analysis of training Error

- Theorem [Freund&Schapire '97]:

  write ε_t as ½-γ_t

  then $\text{training error}(H_{\text{final}}) \le \exp\left(-2\sum_t \gamma_t^2\right)$

  so if $\forall t$: $\gamma_t \ge \gamma > 0$ then

  $$\text{training error}(H_{\text{final}}) \le e^{-2\gamma^2 T}$$

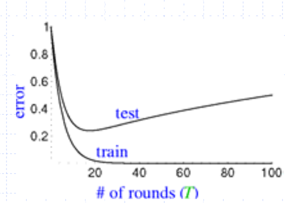
  So, training error continues to drop and reaches

  zero as boosting iteration (T) is increased

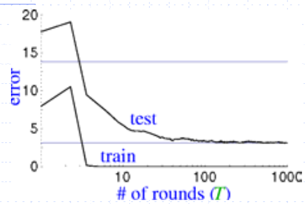
Generalization Error

- Generalization error bound of the final hypothesis in terms
  - training error
  - the sample size
  - VC dimension of the hypothesis space
  - the number of boosting round
- As classifier becomes more complex, test error expected to increase – Occam's razor

Generalization Error Analysis



A Typical Test Run



- Test error does not increase even after 1,000 rounds (~2,000,000 nodes)
- Test error continues to drop after training error is zero!
- Occam's razor wrongly predicts "simpler" rule is better.

Another Argument

- Based on bias and variance – by Breiman and others
- Voting method works by reducing the variance of a learning algorithm
- Useful explanation for bagging but incomplete for boosting
- Large variance not a requirement for boosting

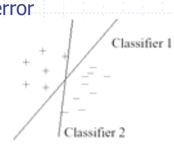
A reasonable argument

" Voting the classifiers does not increase the complexity, but merely *smooth* the prediction"

- The complexity of such combined classifier much greater than the base and may result in overfit

A better Explanation: Margin

- Consider more than just the training error
- Take into account the classifier confidence
- Margin – a measure of classification confidence
- Improvements on margin on the training set guarantees an improvement in the upper bound on the generalization error



Margin

- Boosting constructs hypothesis of the form sgn(f(x))
- The prediction of the combined classifier is the result of the vote over a set of base classifiers. The weights assigned to the base classifiers sums to 1
- The classification margin is defined as the difference between the weight assigned to the correct label and the maximum weight assigned to the incorrect label.

Margin of binary class problem

- For binary class problem, the classification of an example is correct if sgn(f(x)) = y
- So in this case margin = $y \cdot f(x)$

  where

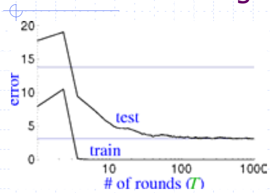
$$f(x) = \frac{\sum_t \alpha_t h_t(x)}{\sum_t \alpha_t} \in [-1,1]$$

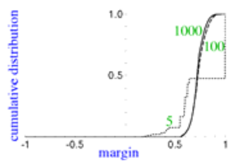
- Higher margin => lower generalization error

Generalization error as a function of margin distribution

- Margin distribution graphs – plot of fraction of examples whose margin is at most x as a function of $x \in [-1, 1]$
- Bagging and Boosting
  - increase the margins associated with training examples
  - converge to a margin distribution with most examples having large margins
- Experiments shows correlation between a reduction between fraction of examples with small margin and improvements in the test error

Effect of boosting on the margin





epoch	5	100	1000
training error	0.0	0.0	0.0
test error	8.4	3.3	3.1
%margins≤0.5	7.7	0.0	0.0
Minimum margin	0.14	0.52	0.55

Bounds on Generalization error

- upper bounds on generalization error of
  - in terms of # training examples
  - complexity of base hypotheses –
  - but not on # of base classifiers
    - considers not only training error but # incorrect classifications, and confidence of classifications
- these bounds imply :
  - # of training examples with small margin drops exponentially fast with the number of base classifiers
- Theorem proves achieving a large margin results in an improved bound

Theorem 1 – finite base-classifier space

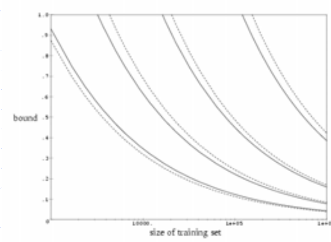
For $\delta > 0$ then with probability at least $1-\delta$ every weighted average function f satisfies the following bound for all $\theta > 0$:

$$\mathbf{P}_{\mathcal{D}}\left[yf(x) \le 0\right] \le \mathbf{P}_S\left[yf(x) \le \theta\right] + O\left(\frac{1}{\sqrt{N}}\left(\frac{\log N \log |\mathcal{H}|}{\theta^2} + \log(1/\delta)\right)^{1/2}\right)$$

Taking N to infinity, and by substituting for # of hypotheses:

$$\mathbf{P}_{\mathcal{D}}\left[yf(x) \le 0\right] \le \mathbf{P}_S\left[yf(x) \le \theta\right] + \sqrt{\frac{2 \ln N \ln |\mathcal{H}|}{N\theta^2}} + o\left(\sqrt{\frac{\ln N}{N}}\right)$$

Bound dependence on training set



Plots of second term of new bound (approximation, dotted line) with second term of old bound (solid line) for theta = 1/20, 1/8, ¼ and ½ (up – down)

Theorem 2 – infinite base-classifier space

Assume N (number of training examples) >= d >= 1

For $\delta > 0$ then with probability at least $1-\delta$ every weighted average function f satisfies the following bound for all $\theta > 0$:

$$\mathbf{P}_{\mathcal{D}}\left[yf(x) \le 0\right] \le \mathbf{P}_S\left[yf(x) \le \theta\right] + O\left(\frac{1}{\sqrt{N}}\left(\frac{d \log^2(N/d)}{\theta^2} + \log(1/\delta)\right)^{1/2}\right)$$

Effect of boosting on margin distribution

Suppose the base learner when called by AdaBoost generates weighted training errors: $\epsilon_1, \ldots, \epsilon_M$ then for any theta we have:

$$\mathbf{P}_{(x,y)\sim S}\left[yf(x) \le \theta\right] \le 2^M \prod_{m=1}^{M} \sqrt{\epsilon_m^{1-\theta}(1-\epsilon_m)^{1+\theta}}$$

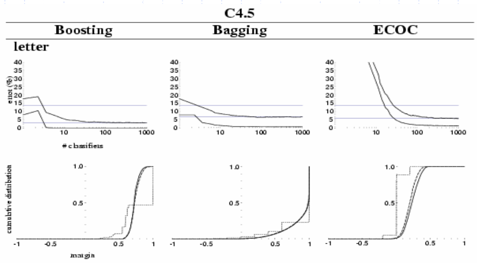
If error<½ for all D_t then the training error of the combined classifier decreases exponentially fast with M.

In effect: the larger our aggregation size M, the more we shift the distribution of margins towards the right.

Example: training error rate is ¼ for all rounds, then for theta = 0:

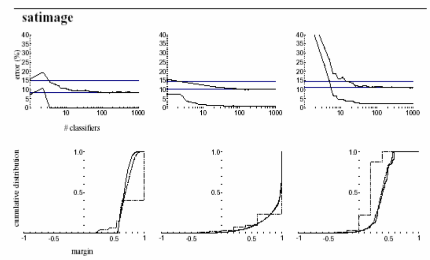
	right hand term
M=10 :	0.2373
M=100:	5.663 x 10^-7

Experimental Results

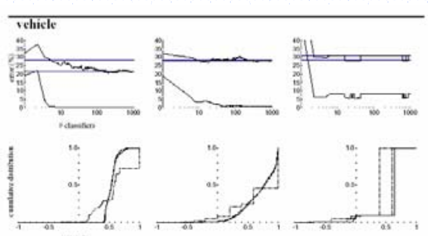


iterations: 5 – short dashed, 100 – long dashed, 1000 solid

Experimental Results



Experimental Results



Relation to Bias – Variance theory

- Bias – Variance Decomposition
  - Separate the expected error of a classifier into a *bias* term and a *variance* term.
  - Bias measures the persistent error of a learning algorithm
  - Variance term measures the error due the fluctuations for one single classifier.

Definition of Bias and Variance

- By Kong & Dietterich

  C_S : classification rule from one base learning given training set S.
  C_A : classification rule from majority vote of base learners, each which are run on infinite # of training sets
  C* : Bayes optimal prediction rule given distribution D.

$PE(C)= P_{(x,y)\sim D}\ [C(x) \neq y]$

$Bias = PE(C_A) - PE(C^*)$

$Variance = E_S{\sim}D^m[PE(C_S)] - PE(C_A)$

Bagging and variance reduction

- Under idealized condition Variance is decrease in error effected by bagging a large number of base classifier
- Ideal situation – bootstrap samples used in bagging faithfully approximate truly independent samples
- Poor performance in reality – ideal condition is not met in practical

Boosting and variance reduction

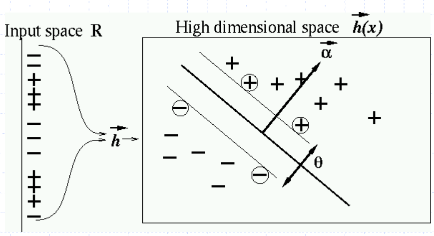
- Boosting a variance reduction procedure – by Breiman
- Experiments shows boosting does more than reducing variance
- Theorem suggest different characterization
  - Poor performance of boosting
    - Insufficient training data
    - Training error become large too quickly

Averaging increases complexity

- Voting seen as smoothing or averaging a simple classification rule (the weak learner)
- There are cases where training error zero
  - But high generalization error.
- This behavior matches overfitting, and is a result of the combined classifier fitting the training set exactly, in this case the complexity of the average rules is too large.
- Margin based analysis gives a correct explanation – the margin is low.

Relation to SVMs

SVM: map x into high-dim space, separate data linearly



Relation to SVM cont….

- Aims to find a linear combination is high dimensional space which has large margin on the instances
- SVM- maximize the margin
- Boosting- minimize an exponential weighting of examples as function of their margin

Conclusion

- A new approach to analyze the generalization error of voted classifier
- Upper bound on prediction error
- Error is a function of empirical distribution of margin
- Boosting finds classifier with large margin
- Open problem: Does there exist a better bound

Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods

Robert E. Schapire Yoav Freund
Peter Bartlett Wee Sun Lee
Presented by : Sweta Sinha
Commentary: Krishna V Chengavalli

Handling Multi Class Problems

- Real World problems are generally multi-class
  - Eg. OCR problem

Some methods to deal them
One Versus Rest
Pair wise classification

Variant of Boosting

- Predict plausible classes
- Combined classifier chooses most frequent label from plausible sets
- Pseudoloss measure
- Overcomes the necessity of having ½ accuracy for base classifiers

Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods

By,
Schapire, Freund, Bartlett and Lee

Comments by,
Srikanth Varanasi

AdaBoost and SVMs - differences

- Different norms can result in very different margins
  1. difference in norms may not be very significant in low dimensional spaces
  2. in high dimension spaces, difference in norms can result in very large margin difference
- When number of relevant weak hypotheses is a small fraction of total weak hypotheses – margin in AdaBoost is larger

Differences cont..

- Computation requirements are different
- Computation involved is maximizing the margin
- SVMs corresponds to *quadratic programming* and AdaBoost corresponds to *linear programming*
- *Quadratic programming* is more computationally demanding

Differences cont..

- A different approach is used to search efficiently in high dimensional space
- SVMs use kernels which allow to perform low dimensional calculations
- AdaBoost employs greedy search