Boosting and Machine Learning

Marc Sebban

Noviembre 2004, Universidad de Alicante
Outline

• Learning and Learnability

• Where BOOSTING comes from?

• AdaBOOST and main theoretical results

• Drawbacks on real world problems (noise, convergence speed)

• Main approaches to overcome these drawbacks

• Current research and interesting perspectives (in my mind!)
Learning and Learnability
(Supervised) Machine Learning

Objective: Construction of a model of the reality from (symbolic and/or numerical) data

→ More formally, identify a hypothesis $\hat{h}_S$, from a learning sample $S$, the “closest” from the optimal hypothesis $h^*$ from among a class of hypotheses $H$, which is able to learn a target concept $f$.

→ Two different classes $H_1$ and $H_2$ lead to different hypotheses $\hat{h}_S^1$ and $\hat{h}_S^2$. 
How can one choose between hypotheses?

Intuitive approach

→ Occam’s Razor principle: choose from a set of otherwise equivalent models of a given phenomenon the simplest one.
Solution for selecting a reliable hypothesis: Bias/Variance Compromise

This compromise expresses the existing tradeoff between the power of a class of hypotheses and its instability in front of data.

The real risk (generalization error) of a given hypothesis $\hat{h}_S$, called $R_{real}(\hat{h}_S)$, is equal to the sum of its bias and its variance.

$$R_{real}(\hat{h}_S) = B(\hat{h}_S) + V(\hat{h}_S)$$

- Bias expresses the degree of fit of $\hat{h}_S$ to learn the target.
- Variance expresses the variability of $\hat{h}_S$ according to $S$. 
Bias/Variance Compromise

$H$

$\hat{h}_S \rightarrow h^*$

variance

bias

Real risk of $h$

$f$
What is the goal of Learning? Select a reliable hypothesis which provides a good compromise between its bias and its variance...

→ In order to learn, one needs an inductive principle. The most used in ML is the ERM inductive principle (also used in Boosting).
ERM Inductive Principle

Principle of the Empirical Risk Minimization: *a hypothesis that works on the learning sample should also work in general.*

(Other inductive principles: Bayesian approach, maximum likelihood, information compression, etc.)

Is it always true? Unfortunately, no... except under convergence constraints.

**More formally:** how can one ensure that an inductive principle (ERM here) results in the decrease of the real risk?
**ERM Inductive Principle**

ERM is valid if and only if the real risk associated to $\hat{h}_S$ (the hypothesis built from $S$ with a minimal empirical risk $R_{emp}$) is ensured to be close to the real risk of the unknown optimal hypothesis $h^*$.

$$\hat{h}_S = \arg\min_{h \in H} R_{emp}(h)$$

$$h^* = \arg\min_{h \in H} R_{real}(h)$$

This must be true for the majority of cases, such that:

$$\forall 0 \leq \epsilon, \delta \leq 1, P(|R_{real}(\hat{h}_S) - R_{real}(h^*)| \geq \epsilon) \leq \delta$$
PAC Model

Intuitively, we can think that the nature of the learning sample $S$ (i.e. its distribution) and its size ($m$) have a direct effect on the correlation between $\hat{h}_S$ and $h^*$.

PAC Model (Probably Approximately Correct), and Valiant’s work (1984) (A Theory of the Learnable), give a theoretical answer to the following problem.

**The problem:** what are the conditions under which $R_{emp}(\hat{h}_S)$ converges toward $R_{real}(h^*)$?
If $|H|$ is finite

It requires the two following conditions:

1. that $R_{emp}(\hat{h}_S)$ converges toward $R_{real}(\hat{h}_S)$

$$P(\max_h |R_{real}(\hat{h}_S) - R_{emp}(\hat{h}_S)| \geq \epsilon) < 2|H|e^{-2\epsilon^2m}$$

$$m > \frac{2}{\epsilon^2 \ln \frac{2|H|}{\delta}}$$

to bound this deviation by $\delta$.

2. that the real risk $R_{real}(\hat{h}_S)$ converges toward $R_{real}(h^*)$

$$P(|R_{real}(h^*) - R_{real}(\hat{h}_S)| \geq \epsilon) < 2e^{-2\epsilon^2m}$$
If $|H|$ is infinite

We use the Vapnik-Chervonenkis dimension which provides an idea about the power of $H$

*The Vapnik Chervonenkis dimension $d_H$ is defined to be the maximum number of examples that can be shattered by $H$.*

*A set of examples is shattered by $H$ if $H$ is able to express all the dichotomies of $S$.*

Then, the VC-Dim is the maximal number of examples for which $H$ is able to learn any labelling.
Bounds on the size of the learning sample

\[ m = O\left(\frac{1}{\epsilon} \ln\frac{1}{\delta} + \frac{d_H}{\epsilon} \ln\frac{1}{\epsilon}\right) \]

For example, using linear separators, with a deviation \( \epsilon = 0.01 \) and a confidence threshold of 95\% (\( \delta = 0.05 \)), then \( m \geq 2425 \ldots \)

Many bounds on the real risk have been proposed, often under the following form:

\[ R_{real}(h) \leq R_{emp}(h) + \phi\left(\frac{d_H}{m}\right) \]
To sum up...

To efficiently bound the real risk, one must either:

- know the penalization term $\phi \left( \frac{d_H}{m} \right)$ to correct the too optimistic estimation of the empirical risk.

$$R_{real}(h) \leq R_{emp}(h) + \phi \left( \frac{d_H}{m} \right)$$

But, is it possible to have a high number $m$ of learning examples to decrease this bound?

- or, when the VC-Dim is unknown, correctly estimate with $\hat{\theta}$, without any bias and with a small variance, the real risk.

$$E(\hat{\theta}) = R_{real}(h)$$
Main empirical estimation methods

- Estimation using a test sample
- Estimation using a cross-validation procedure
- Estimation using bootstrap
- Estimation using classifier combination: Bagging et Boosting.
Boosting Theory

Yoav Freund and Robert Schapire

"A Decision Theoretic Generalization of On-Line Learning and an Application to Boosting"

Intuitive approach...

A horse-racing gambler wants to maximise his winnings (Freund & Schapire 1999) with the help of an expert.

Two problems to solve:

- How should one choose the races presented to the expert?
- How should one combine the rules into a single, highly accurate prediction rule?

→ **Bagging**: races are randomly chosen and the final hypothesis is a unweighted classifier combination.

→ **Boosting**: races are chosen according to their difficulty to be learnt, and the final hypothesis is a weighted classifier combination.
Boosting Theory

Schapire proposed the first boosting algorithm (in *The strength of weak learnability* 1990) to answer to the following question:

"Tell me Rob, is it possible to obtain a strong hypothesis, from a PAC point of view, by using only a weak learner, i.e. which is just a little bit better than a random guess?"

→ Birth of ADABOOST (which deals with binary problems)
AdaBoost

Input \( S = \{(x_1, y_1), \ldots, (x_{|S|}, y_{|S|})\} \)

for all \( x \in S \) do \( w_0(x) \leftarrow 1/|S|; \)

for \( t = 0 \) to \( T \) do

\[ h_t \leftarrow \text{WL}(S, w_t); \]
\[ \epsilon_t \leftarrow \sum_{x \in S: y \neq h_t(x)} w_t(x); \]
\[ \alpha_t \leftarrow \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right); \]
\[ Z_t \leftarrow \sum_{x \in S} w_t(x) \exp(-\alpha_t y h_t(x)); \]

for all \( x \in S \) do

\[ w_{t+1}(x) \leftarrow w_t(x) \exp(-\alpha_t y h_t(x))/Z_t; \]

\[ f(x) = \sum_{t=0}^{T} \alpha_t h_t(x); \]

return \( H_T \) such that \( H_T(x) = \text{sign}(f(x)) \);
Theoretical Results in Learning (1)

**Theorem 1**: Bound on the empirical risk (learning error)

\[
R_{emp} = \frac{1}{m}(\{i / H(x_i) \neq y_i\}) \leq \frac{1}{m} \sum_i \exp(-y_i f(x_i)) = \prod_t Z_t
\]

This theorem means that for minimizing the empirical risk, one must minimize the product of the \(Z_t\), and then *a fortiori* each \(Z_t\).
Theoretical Results in Learning (2)

**Theorem 2:** For minimizing $Z_t$, the confidence coefficient of each hypothesis $\alpha_t$ must be equal to:

$$\alpha_t = \frac{1}{2} \log \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)$$

Simple optimization problem with partial derivatives
Theoretical Results in Learning (3)

**Theorem 3**: Exponential decrease of the empirical error

\[
\prod_{t=0}^{T} (Z_t) = \prod_{t=0}^{T} (2\sqrt{\varepsilon_t (1 - \varepsilon_t)}) = \prod_{t=0}^{T} \sqrt{1 - 4\gamma_t^2} < \exp (-2 \sum_{t=0}^{T} \gamma_t^2)
\]

where \(\varepsilon_t = \frac{1}{2} - \gamma_t\)

\(\gamma_t\) expresses the “just a little bit better than a random guess”

In other words, this theorem means that the empirical risk decreases with the number of iterations \(T\).

But, what about the generalization error?
Theoretical Results in Generalization: margin maximization (1)

The margin is defined to be the minimal distance between any training example $x$ and the separating hyperplane $f(x)$.

$$margin(x) = y \sum_{t} \alpha_t h_t(x) = yf(x)$$

Schapire et al. have proven that wide margins over $S$ have a direct impact on the real risk $R_{\text{real}}(h)$. More formally,

$$R_{\text{real}}(h) \leq Pr(margin(x) \leq \theta) + O(\sqrt{\frac{d_h}{m\theta^2}})$$
Theoretical Results in Generalization: margin maximization (2)

Boosting is able to increase the margin with T, i.e. decrease $Pr(margin(x) \leq \theta)$

**Theorem 4:**
For any $\theta$:

$$P(yf(x) \leq \theta) \leq (\sqrt{(1 - 2\gamma)^{1-\theta}(1 + 2\gamma)^{1+\theta}})^T$$

This theorem means that if $\theta < \gamma$ the expression inside the parentheses is smaller than 1 so that the probability that $yf(x) \leq \theta$ decreases exponentially fast with $T$. 
Summary of boosting properties

- The empirical risk exponentially decreases toward 0 with the number of iterations $T$.
- The real risk also exponentially decreases even when the empirical risk has reached 0.
- The classifier combination allows us to decrease the variance of the real risk, whatever the algorithm we use.
In practice, AdaBoost often performs well... :-)

but... :-(

Drawbacks of Boosting on real world problems

- Noisy data
- Non optimal representation spaces → Overlaps between probability distributions
Overfitting in the presence of noise

Learning problem with 2 classes and 5% of noise

![Diagram showing overfitting in the presence of noise]
Solutions for overcoming such a problem

Since the uncontrolled increase of some examples explains in part the overfitting phenomenon, 2 solutions are available:

- Use of other weighting rules:
  - MADABOOST (Domingo & Watanabe, COLT 2000),
  - BROWNBOOST (Freund, Machine Learning 2001),
  - GENTLE ADABOOST (Friedman et al., TR 1998)
  - Soft Margins (Ratsch et al., NIPS 1998)
  - iADABOOST (Sebban & Suchier, ECML 2003)

- Detect and remove a priori noisy data, outliers:
Convergence Problem

Learning problem with a Bayesian error of 0.2
Solution to the problem of convergence
(Nock & Sebban, PRL 2000)

**Theorem 4**: Using ADABOOST, the following bound on the empirical bound $H$ holds:

$$\frac{|\{i : H(x) \neq y_i\}|}{m} \leq \left( \prod_{t} Z_t \right) E_{D_{T+1}}[\delta(x)] + \epsilon^*$$

where $\epsilon^*$ is the minimal error on $S$ and $E_{D_{T+1}}[\delta(x)]$ is the expectation of $\delta(x)$, with

$$\delta(x) = \frac{|n^+(x) - n^-(x)|}{(n^+(x) + n^-(x))}$$
Current trend and Research Perspectives

- Boosting and data reduction techniques (dimensionality reduction, prototype selection)
- New (smaller) bounds, particularly in the presence of noise
- Boosting and co-training/meta-learning
- Boosting and grammatical inference
  - In regular grammatical inference (Sebban & Janodet, ICML 2003).
  - In stochastic grammatical inference, it seems to be difficult... (what is an error? what is the label $y$?) (Thollard & al., ECML 2002).
The End