Boosting as a Product of Experts

Narayanan U. Edakunni, Gavin Brown and Tim Kovacs
University of Bristol and University of Manchester, UK

Introduction

- We establish a **generic probabilistic model** of boosting.
- The class probability is represented as a **Product of Experts** (PoE).
- Boosting is modeled as addition of new experts to the PoE with the addition of an expert increasing the likelihood at each iteration.
- Condition that the likelihood does not decrease with addition of an expert leads to the **weak learning** criterion.
- We show that for a specific parameterization of the expert probabilities, **PoE = AdaBoost**.
- Using a specific form of expert probability, POEBoost.CS is developed, which uses base hypotheses with **probabilistic outputs**.
- The comparisons of POEBoost.CS with existing boosting algorithm illustrates the advantages of having a probabilistic model for boosting.

Product of Experts [Hinton, 2002]

The conditional class probability of label y given input x is given by

$$P(Y=y|X=x,h_1\dots h_M) = \frac{\prod_{j=1}^{M} P(Y=y|X=x,h_j)}{\prod_{j=1}^{M} P(Y=y|X=x,h_j) + \prod_{j=1}^{M} P(Y=\overline{y}|X=x,h_j)}$$

where Y and X are random variables for output and input, $\overline{y}$ is the complement of y and h the hypothesis.

Incremental expansion of PoE

Addition of an expert $P(y|x,h_j)$ to a PoE P_{j-1} with $j-1$ experts gives

$$P_j(y|x) = \frac{P(y|x,h_j)P_{j-1}(y|x)}{P(y|x,h_j)P_{j-1}(y|x) + P(\overline{y}|x,h_j)P_{j-1}(\overline{y}|x)} \quad (1)$$

where P_j is a PoE with j experts.

Monotonic increase in likelihood

- The log likelihood is
$$\mathcal{L}_j = \log P(y_1\dots y_N|x_1\dots x_N, h_1\dots h_j)$$
where the data points $(x_1,y_1)\dots(x_N,y_N)$ are assumed to be IID.
- With every addition of an expert, the likelihood does not decrease if
$$\Delta\mathcal{L} = \mathcal{L}_j - \mathcal{L}_{j-1} \geq 0.$$
It leads to a constraint on the new expert
$$\sum_{i=1}^{N} \frac{D_i^{j-1}}{P(y_i|x_i,h_j)} \leq 2 \quad (2)$$
where $D_i^{j-1} = \frac{P(\overline{y}_i|x_i,h_1\dots h_{j-1})}{\sum_{i=1}^{N} P(\overline{y}_i|x_i,h_1\dots h_{j-1})}$ such that $\sum_i D_i = 1$.
- We can now come up with a procedure to expand the PoE incrementally using eq. (1) to add new hypotheses, with each hypothesis satisfying the constraint given in eq. (2).

Algorithm : Incremental learning in PoE

Training data : $S = ((x_1,y_1)\dots(x_N,y_N))$
$D_i^1 = 1/N$ for $i = 1\dots N$.
for j = 1 to M **do**
Obtain a hypothesis h_j such that
$\sum_{i=1}^{N} \frac{D_i^j}{P(y_i|x_i,h_j)} \leq 2$
$D_i^{j+1} = P(\overline{y}_i|x_i,h_j)D_i^j$
$D_i^{j+1} = \frac{D_i^{j+1}}{\sum_i D_i^{j+1}}$
end for

POEBoost.DS : Discrete Symmetric expert

- We assume a certain form for the expert probability.
  - Parametrize expert probability
  $$P(Y=y|X=x,h) = P(Y=y|Z=y)P(Z=y|x,h) + P(Y=y|Z=\overline{y})P(Z=\overline{y}|x,h)$$
  - Symmetric class error probability
  $$P(Y=-1|Z=1) = P(Y=1|Z=-1) = P_e$$
  - Discrete class error probability
  $$P(Z=y|x,h), P(Z=\overline{y}|x,h) \in \{0,1\}$$
  - Error probability parameterization
  $$P_e = \frac{e^{-\alpha}}{(e^{-\alpha}+e^{\alpha})}$$
- We show that these assumptions on the expert reduces incremental learning in a PoE to AdaBoost.

Product of Experts = majority voting

Conditional class probability is given by

$$P(Y=y|X=x,h_1\dots h_M) = \frac{e^{y\sum_j \alpha_j h_j(x)}}{e^{y\sum_j \alpha_j h_j(x)} + e^{-y\sum_j \alpha_j h_j(x)}}$$

The class y_q for an input x_q is given by

$$y_q = sign\left(\sum_j \alpha_j h_j(x_q)\right)$$

equivalent to prediction rule of AdaBoost [Freund and Schapire, 1997].

Incremental expansion of PoE = updating weights

The weights on the data points D_i^{j-1} is updated with the addition of an expert h_j with voting weight α_j as

$$D_i^j \propto e^{-y_i\sum_j \alpha_j h_j(x_i)} D_i^{j-1}$$

similar to that of the AdaBoost algorithm.

Monotonic increase in likelihood = weak learning condition

- The likelihood does not decrease with the addition of an expert if
$$\epsilon_j \leq P_e^j \leq \frac{1}{2}$$
where $\epsilon_j = \sum_{i:h_j(x_i)\neq y_i} D_i^{j-1}$.
- The likelihood constraint on h_j equivalent to that of weak learning condition of AdaBoost [Freund and Schapire, 1997, Schapire, 1990].
$$\epsilon_j \leq \frac{1}{2}$$
- The update rule for α_j
$$0 \leq \alpha_j \leq \frac{1}{2}\log\frac{1-\epsilon_j}{\epsilon_j}$$
AdaBoost uses $\frac{1}{2}\log\frac{1-\epsilon_j}{\epsilon_j}$ for α_j.

POEBoost.CS : Continuous Symmetric expert

- We now slightly adapt the assumptions of POEBoost.DS to obtain a variant.
  - Symmetric class error probability
  $$P(Y=-1|Z=1) = P(Y=1|Z=-1) = P_e$$
  - Continuous class error probability
  $$P(Z=y|x,h), P(Z=\overline{y}|x,h) \in [0,1]$$
- We can now derive an algorithm that can handle base hypotheses with probabilistic class predictions, without a large modification in the algorithm.

Weak learning condition

The likelihood does not decrease with the addition of an expert if

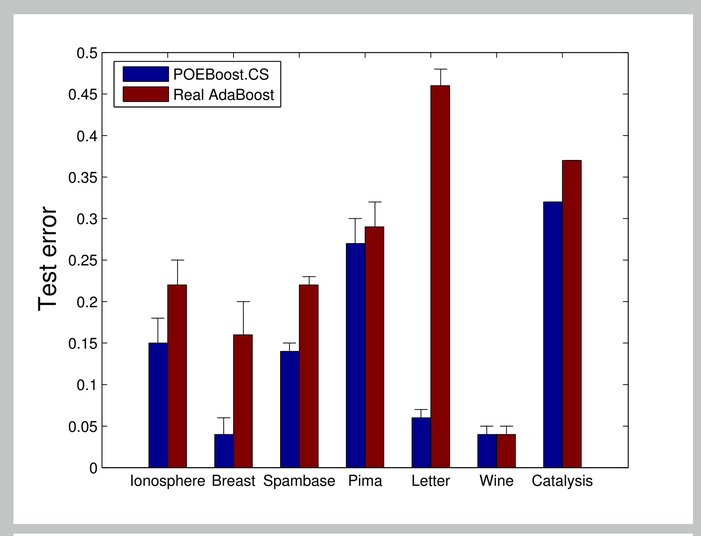
$$\epsilon_j^c \leq P_e^j \leq \frac{1}{2}$$

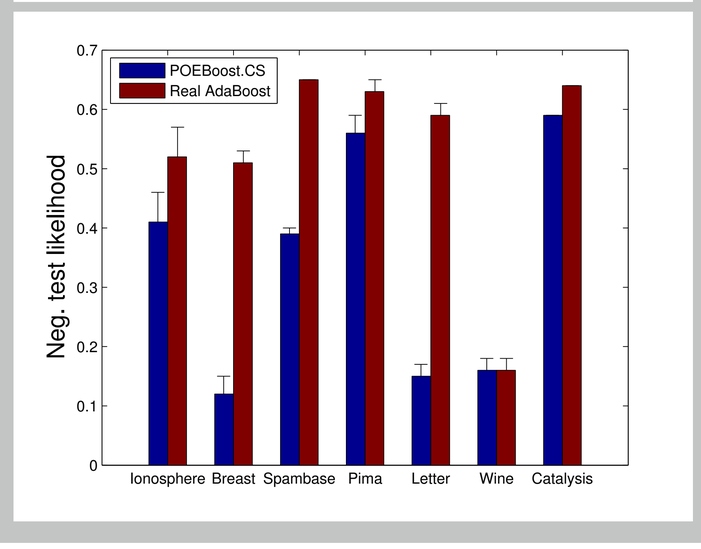
where

$$\epsilon_j^c = \left[\sum_{i:P(z=y_i|x_i,h_j)\leq 0.5} D_i^{j-1}(2P(Z=y_i|x_i,h_j)-1)\right] / \left[2\sum_{i:P(z=y_i|x_i,h_j)\leq 0.5} D_i^{j-1}(P(Z=y_i|x_i,h_j)-1) - 2\sum_{i:P(z=y_i|x_i,h_j)\geq 0.5} D_i^{j-1}P(Z=y_i|x_i,h_j)+1\right].$$

Evaluations - POEBoost.CS vs Real AdaBoost [Schapire and Singer, 1998]

- Comparison of ability to generalize on real world datasets.





Conclusions

- We have presented a novel interpretation of boosting algorithms as a PoE.
- The sequential distribution updates in the boosting explained as an iterative model selection process.
- The model we have presented can incorporate arbitrary probabilistic models as the experts, including real-valued predictions naturally.
- In a particular special case, the model reduces to an AdaBoost-like algorithm.

References

- Freund, Y. and Schapire, R. E. (1997).
A decision-theoretic generalization of on-line learning and an application to boosting.
Journal of Computer and System Sciences, 55:119–139.
- Hinton, G. (2002).
Training products of experts by minimizing contrastive divergence.
Neural Computation, 14:1771–1800.
- Schapire, R. E. (1990).
The strength of weak learnability.
Machine Learning, 5:197–227.
- Schapire, R. E. and Singer, Y. (1998).
Improved boosting algorithms using confidence-rated predictions.
In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*.

http://www.cs.bris.ac.uk/~nara
nara@cs.bris.ac.uk