Averaged Boosting: A Noise-Robust Ensemble Method*

Yongdai Kim

Department of Statistics, Ewha Womans University
Seoul, Korea, 120-750
ydkim@mm.ewha.ac.kr
http://home.ewha.ac.kr/ ydkim/

Abstract. A new noise robust ensemble method called "Averaged Boosting (A-Boosting)" is proposed. Using the hypothetical ensemble algorithm in Hilbert space, we explain that A-Boosting can be understood as a method of constructing a sequence of hypotheses and coefficients such that the average of the product of the base hypotheses and coefficients converges to the desirable function. Empirical studies showed that A-Boosting outperforms Bagging for low noise cases and is more robust than AdaBoost to label noise.

1 Introduction

Bagging (Breiman [2]) and AdaBoost (Freund and Schapire [7]) are two most commonly used ensemble algorithms. A number of empirical studies including those by Quinlan [10], Bauer and Kohavi [1], Optiz and Maclin [9] and Ditterich [4] indicated that although AdaBoost is more accurate than Bagging in most cases, AdaBoost may overfit high noise cases, thus decreasing its performance.

In this paper, we propose a new noise-robust ensemble method, named "Averaged Boosting (A-Boosting)," which combines the advantages of both the Bagging and AdaBoost method; that is, A-Boosting improves Bagging for low-noise cases and outperforms AdaBoost for high-noise cases.

In section 2, the A-Boosting algorithm is presented. Using a hypothetical ensemble algorithm, we show that A-Boosting can be understood as an algorithm that generates a sequence of ensemble models converging to the optimal function. In section 3, numerical results on several data sets are presented. The paper concludes with a brief discussion in section 4.

2 A-Boosting

2.1 The Algorithm

The A-Boosting algorithm is given as follows.

* This research is supported in part by US Air Force Research Grant F62562-02-P-0547 and in part by KOSEF through Statistical Research Center for Complex Systems at Seoul National University.

K.-Y. Whang, J. Jeon, K. Shim, J. Srivatava (Eds.): PAKDD 2003, LNAI 2637, pp. 388–393, 2003.
© Springer-Verlag Berlin Heidelberg 2003

Algorithm 2. A-Boosting

1. Start with weights $w_i = 1/n, \ \ i = 1, \ldots, n$.
2. Repeat for $m = 1, \ldots, M$;
   a) Fit a hypothesis f_m on $\mathcal{F}$ minimizing the weighted misclassification error rate with weights $\{w_i\}$ on the training examples.
   b) Calculate the misclassification error ϵ_m of f_m on the original training examples by
   $$\epsilon_m = \sum_{i=1}^{n} I(y_i \neq f_m(\mathbf{x}_i))/n.$$
   c) Let $\beta_m = \log((1 - \epsilon_m)/\epsilon_m)/2$
   d) Update $w_i = \exp(-y_i \sum_{t=1}^{m} \beta_t f_t(\mathbf{x}_i)/m), \ i = 1, \ldots, n$ and renormalize so that $\sum_i w_i = 1$.
3. Output the classifier $\text{sign}(\sum_{m=1}^{M} \beta_m f_m(\mathbf{x}))$.

There are two key differences between the A-Boosting and AdaBoost algorithms. The first is that the A-Boosting algorithm uses the average of the product of the base hypotheses and coefficients while AdaBoost uses the sum of it. The second difference is that the A-boost algorithm calculates the coefficient based on the error rate of the current hypothesis on the original training example while the AdaBoost algorithm uses the updated weights. The A-Boosting algorithm in Hilbert space given in the next subsection explains why such modifications are necessary.

2.2 A-Boosting in Hilbert Space

Consider a probability measure Q on R^p defined by

$$Q(d\mathbf{x}) = \frac{\sqrt{p(\mathbf{x})(1 - p(\mathbf{x}))}P(d\mathbf{x})}{\int_{R^p} \sqrt{p(\mathbf{x})(1 - p(\mathbf{x}))}P(d\mathbf{x})}$$

where $p(\mathbf{x}) = \Pr(Y = 1|\mathbf{X} = \mathbf{x})$ and $P(d\mathbf{x})$ is the marginal distribution of $\mathbf{X}$. Let $\mathcal{F}$ be the set of functions in $L_2(Q)$ with norm 1 and let f^* be the target function to be estimated. Assume that $\|f^*\| < \infty$. The A-Boosting algorithm in Hilbert space is given as follows.

Algorithm 3. A-Boosting in Hibert space

1. Set $f_0(\mathbf{x}) \equiv 0$ and $\beta_0 = 0$.
2. Repeat $m = 1, \ldots, M$
   a) Set f_m by
   $$f_m = \text{argmin}_{f \in \mathcal{F}} \left\| f^* - \sum_{t=0}^{m-1} \beta_t f_t/(m-1) - f \right\|.$$

b) Find β_m;

$$\beta_m = \text{argmin}_{a \in R} \|f^* - a f_m\|.$$

3. Estimate $f^*(\mathbf{x})$ by $\sum_{t=1}^{M} \beta_t f_t(\mathbf{x})/M$.

The following theorem proves the convergence of the above algorithm.

Theorem 1. *In Algorithm 3,*

$$\left\| \sum_{t=1}^{M} \beta_t f_t(\mathbf{x})/M - f^*(\mathbf{x}) \right\| \to 0$$

as $M \to \infty$.

Now, we turn to the classfication problem. Consider the population exponential criteria $C(f) = \text{E}(\exp(-Yf(\mathbf{X})))$ for given function f. Let $p(\mathbf{x}) = \Pr(Y = 1|\mathbf{X} = \mathbf{x})$ and let $f^*(\mathbf{x}) = \log(p(\mathbf{x})/(1-p(\mathbf{x}))/2$. Lemma 1 in Friedman et al. [5] proves that $C(f)$ attains its minimum at $f = f^*$.

Taylor expansion yields

$$\begin{aligned} C(f) &= \int_{R^p} \left(\exp(f(\mathbf{x}) - f^*(\mathbf{x})) + \exp(f^*(\mathbf{x}) - f(\mathbf{x}))\right) Q(dx) \\ &= \int_{R^p} 2 + (f^*(\mathbf{x}) - f(\mathbf{x}))^2 + o((f^*(\mathbf{x}) - f(\mathbf{x}))^2) Q(dx) \qquad (1) \\ &\approx 2Q(R^p) + C\|F - f\|^2 \end{aligned}$$

if we ignore the $o((f^*(\mathbf{x}) - f(\mathbf{x}))^2)$ term. Hence, the population exponential criterion is approximately the same as the square of $L_2(Q)$ norm, and hence the minimizer of $\|f^* - f\|$ on $f \in \mathcal{F}$ is approximately the same as the minimizer of the population exponential criterion $C(f)$. Using this property, the modification of Algorithm 3 with the population exponential criterion becomes

Algorithm 4. A-Boosting with the population exponential criterion

---

1. Set $f_0(\mathbf{x}) \equiv 0$ and $\beta_0 = 0$.
2. Repeat $m = 1, \ldots, M$
   a) Set f_m by

$$f_m = \text{argmin}_{f \in \mathcal{F}} C\left(\sum_{t=0}^{m-1} \beta_t f_t + f\right).$$

   b) Find β_m;

$$\beta_m = \text{argmin}_{a \in R} C(a f_m).$$

3. Estimate $f^*(\mathbf{x})$ by $\sum_{t=1}^{M} \beta_t f_t(\mathbf{x})/M$.

Now, the A-Boosting algorithm is obtained by replacing the population exponential criteria $C(f)$ by the exponential criteria $C_n(f)$ and letting $\mathcal{F}$ be the set of base hypotheses. Note that the norm of any base hypothesis is 1.

3 Empirical Studies

For base hypotheses, unpruned trees are used in A-Boosting and Bagging. For AdaBoost, the optimal tree size is selected using the validation set a priori. For the final decision, 300 base hypotheses are combined. We analyzed 14 benchmark datasets from the UC-Irvine machine learning archive. For datasets without test samples, the generalization errors were calculated by 10 repetitions of 10-fold cross validation.

Table 1 presents the generalization errors. A-Boosting improved Bagging in most cases though AdaBoost had the lowest level of generalization errors in general. However, note that AdaBoost used optimally selected tree sizes while nothing was tuned a priori for A-Boosting. Hence, the degree of superiority of AdaBoost over A-Boosting should be discounted.

Table 2 presents the generalization errors when 10% label noise was added. A-Boosting outperformed AdaBoost in many cases (9 out of 14 cases) and was comparable to Bagging. Also, the increases in error rates due to the label noise indcate that A-Boosting is much more robust to label noise that AdaBoost. Bagging is the most robust.

In summary, our empirical results confirm that (i) A-Boosting improves Bagging in most cases, (ii) A-Boosting is much more robust to label noise than AdaBoost, and (iii) A-Boosting is comparable to Bagging for high noise cases.

In practice, non-experts in ensemble methods can use A-Boosting, as almost nothing needs to be tuned a priori. In contrast, AdaBoost requires tuning to find both the optimal number of iterations in addition to the optimal complexity of base hypotheses a priori.

Table 1. Generalization errors

ID	Data Set	A-Boosting	AdaBoost	Bagging
1	Breast Cancer	0.0348	0.0322	0.0332
2	Pima-Indian-Diabetes	0.2400	0.2598	0.2415
3	German	0.2328	0.2481	0.2351
4	Glass	0.2261	0.2204	0.2333
5	House-vote-84	0.0489	0.0439	0.0467
6	Image	0.0566	0.0485	0.0638
7	Ionosphere	0.0698	0.0597	0.0802
8	kr-vs-kp	0.0031	0.0034	0.0075
9	Letter	0.0465	0.0252	0.0640
10	Satimage	0.0880	0.0775	0.0975
11	Sonar	0.1615	0.1180	0.1825
12	Vehicle	0.2411	0.2107	0.2516
13	Vowel	0.4588	0.5844	0.4740
14	Waveform	0.1842	0.1826	0.1842

Table 2. Generalization errors with 10% label noise (Increases of error rates due to label noise, %)

ID	Data Set	A-Boosting	AdaBoost	Bagging
1	Breast Cancer	0.1300 (273.56)	0.1483 (360.55)	0.1279 (285.24)
2	Pima-Indian-Diabetes	0.3024 (26.00)	0.3103 (19.43)	0.3010 (24.63)
3	German	0.3026 (29.98)	0.3291 (32.64)	0.2995 (27.39)
4	Glass	0.3147 (39.18)	0.3047 (38.24)	0.3109 (33.26)
5	House-vote-84	0.1433 (193.04)	0.1818 (314.12)	0.1402 (200.21)
6	Image	0.1402 (147.70)	0.1761 (263.03)	0.1404 (120.06)
7	Ionosphere	0.1685 (141.40)	0.1837 (207.70)	0.1657 (106.60)
8	kr-vs-kp	0.1159 (3638.70)	0.1065 (3032.35)	0.1228 (1537.33)
9	Letter	0.1525 (227.95)	0.1580 (526.98)	0.1505 (135.15)
10	Satimage	0.1575 (78.97)	0.1520 (96.12)	0.1645 (68.71)
11	Sonar	0.2575 (59.44)	0.2305 (95.33)	0.2660 (45.75)
12	Vehicle	0.2905 (20.48)	0.2786 (32.22)	0.2965 (17.84)
13	Vowel	0.5303 (15.58)	0.5497 (-5.93)	0.5368 (13.24)
14	Waveform	0.2398 (30.18)	0.2500 (36.91)	0.2464 (33.76)

4 Discussions

In this paper, we proposed a noise-robust ensemble method called A-Boosting. Using the hypothetical ensemble algorithm in Hilbert space, we explained that A-Boosting can be understood as a method of constructing a sequence of ensemble models converging to the optimal model. Empirical results showed than A-Boosting outperforms Bagging for low noise cases and is much more robust to label noise than AdaBoost.

Most explanations of AdaBoost are related to the numerical optimization algorithms of a certain cost functions (Schapire and Singer [13], Friedman et al. [5], Mason et al. [8], Friedman [6]). One exception is Breiman [3], who conjectures that the success of AdaBoost is its ability to generate an ergodic sequence of base hypotheses whose average converges to the optimal function. A-Boosting can be considered as a deterministic ensemble algorithm of implementing Breiman's idea.

References

1. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, Boosting and Variants. Machine Learning. **36** (1999) 105–139.
2. Breiman, L.: Bagging predictors. Machine Learning. **24** (1996) 123–140.
3. Breiman, L.: Random forests. Machine Learning. **45** (2001) 5–32.
4. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting and Randomization. Machine Learning, **40** (2000) 139–157.
5. Friedman, J.H., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics. **38** (2000) 337–374.

6. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of Statistics. **39** (2001) 1189–1232.
7. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. **55** (1997) 119–139.
8. Mason, L., Baxter, J., Bartlett, P.L. and Frean, M.: Functional gradient techniques for combining hypotheses, In: A.J. Smola, P. Bartlett, B. Schölkopf and C. Shuurmans (Eds.): Advances in Large Margin Classifiers, Cambridge, MA: MIT Press (2000).
9. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research. **11** (1999) 169–198.
10. Quinlan, J.: Boosting first-order learning. In: Arikawa S., Sharma (Eds.): LNAI, Vol. 1160: Proceedings of the 7th International Workshop on Algorithmic Learning Theory, Berlin Springer (1996) 143–155.
11. Räsch, G., Onoda, T., Müller, K.R.: Soft margins for AdaBoost. Machine Learning. **42** (2001) 287–320.
12. Schapire, R., Freund, Y., Bartlett, P., Lee, W.: Boosting the margin: a new explanation for the effectiveness of voting methods. Annals of Statistics. **26** (1998) 1651–1686.
13. Schapire, R., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning. **37** (1999) 297–336.

A Appendix: The Proof of Theorem 1

We will prove that

$$\|f^* - H_m\| \leq \frac{2\|f^*\|}{\sqrt{m}} \tag{2}$$

by use of the mathematical induction. For $m = 1$, (2) is trivially true since $\|f^* - \beta_m f_m\| \leq \|f^*\|$. Suppose (2) holds for $m > 1$. Then, we have

$$\begin{aligned}
\|f^* - H_{m+1}\|^2 &= \left\| \frac{m}{m+1}(f^* - H_m) + \frac{1}{m+1}(f^* - \beta_{m+1} f_{m+1}) \right\|^2 \\
&= \frac{m^2}{(m+1)^2}\|f^* - H_m\|^2 + \frac{1}{(m+1)^2}\|f^* - \beta_{m+1} f_{m+1}\|^2 \\
&\leq \frac{(2\|f^*\|)^2}{m+1},
\end{aligned}$$

since $< f^* - H_m, f^* - \beta_{m+1} f_{m+1} >= 0$. □