BOOSTING THE MARGIN
PAPER REVIEW

Minhaj Fahad
Department of Computer Science
Cornell University
msf257@cornell.edu

Eddie Ramirez Saquic
Department of Computer Science
Cornell University
ebr66@cornell.edu

Oluwasola Ogundare
Department of Operations Research
Cornell University
odo5@cornell.edu

May 11, 2024

ABSTRACT

One of the surprising recurring phenomena observed in experiments with boosting is that the test error of the generated classifier usually does not increase as its size becomes very large, and often is observed to decrease even after the training error reaches zero. In "Boosting the Margin : A New Explanation for the Effectiveness of Voting Method", the authors relate this phenomenon to the distribution of *margins* of the training examples with respect to the generated voting classification rule, where the margin of an example is simply the difference between the number of correct votes and the maximum number of votes received by any incorrect label. The authors show that techniques used in the analysis of Vapnik's support vector classifiers and of neural networks with small weights can be applied to voting methods to relate the margin distribution to the test error. We analyze the efficacy and short-comings of the ideas presented in the paper, as well as contribute new theoretical bounds to classifier differences and show that the phenomena observed in boosting generalization persist in modern classification problems.

Keywords Boosting · Bootstrap · Marginal Distribution · Base-Classifier.

1 Summary

"Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods" by Schapire *et al.* [1] presents a technical exploration of why boosting algorithms continue to perform well even as the complexity of the generated classifier increases significantly. Their investigation reveals that the phenomenon can be explained by the distribution of margins, a theory introduced by Schapire *et al.* [1].

The margin of a training point (x_i, y_i) is a number $\theta \in [-1, 1]$, which can be loosely interpreted as the classifier's confidence of it's prediction for that point. Formally, we say that $f(x) = \sum_j \alpha_j h_j(x)$ is a *voting classifier* if $\alpha_j \geq 0$ for all j. Note that one can additionally assume without loss of generality that $\sum_j \alpha_j = 1$, since normalizing each α_i by $\sum_j \alpha_j$ leaves the sign of $f(x_i)$ unchanged. The margin of a point (x_i, y_i) with respect to a voting classifier f is then defined as

$$\text{margin}(x_i) := y_i f(x_i) = y_i \sum_j \alpha_j h_j(x_i).$$

Moreover, their discussion establishes "rigorous" bounds, defined over finite and infinite base classifier spaces, on the generalization error that relate to the margin distribution, independent of the number of base classifiers used. This provides a novel understanding that has challenged traditional beliefs about classifier complexity and generalization error, suggesting that the key to the success of boosting methods lies in their ability to manipulate the margin distribution rather than merely reducing bias or variance without over fitting.

After presenting experimental evidence which roughly supports the theorems presented, the paper concludes by recognizing that their bounds seem to explain the experiments qualitatively, while their quantitative predictions are

greatly overpessimistic. Finding better, more tight, bounds and bounds that are functions of other statistics is left as an open question to the reader.

2 Advocate Review

The paper serves as a turning point in the understanding of boosting algorithms for classification problems. At the time most machine learning classification research focused on just training and/or testing error, but Schapire *et al.* [1] bring attention to the *confidence* of such classifiers, formally through the idea of margin theory. Despite it's elaborate nature, the paper effectively introduces the concept in a simple and intuitive way, allowing the reader to grasp why it is important and can hold up in practice. Furthermore, the paper is able to successfully continue this approach of introducing the most important topics in a similar manner. Although it does include more technically involved concepts, readers from diverse academic backgrounds are able to grasp the most important revelations and understand how the theorems proved can serve as guiding principles for future practices in the field.

Strengths

One important piece of this paper is it's experimental support for it's theoretical arguments. The explained implementation of Adaboost for classification tasks on popular datasets at the time, in conjunction with informative figures, lends clarity to the claims and theorems, making them comprehensible and accessible to the reader.

In the first experiment, Breiman's bagging method was applied to the C4.5 algorithm. The experiment involved rerunning C4.5 multiple times on different bootstrap subsamples of the training data, then combining the resulting trees through a voting mechanism. The figures illustrating this process clearly highlight the improvement in prediction accuracy, demonstrating that the test error of a single run of C4.5 on the dataset was initially 13.8%. However, after combining 1000 trees through bagging, the test error was reduced to 6.6%. The graphical representations make this improvement immediately apparent, providing the reader with a visual understanding of how bagging enhances model accuracy.

In the second experiment, Freund and Schapire's AdaBoost algorithm was employed with C4.5 as the base learning algorithm. Similar to bagging, AdaBoost runs C4.5 multiple times but selects training subsamples differently, focusing on the most challenging examples by adjusting the distribution of data based on the previous classifier's performance. Informative figures vividly depict how training error drops to zero after a few rounds, while test error continues to decrease with additional rounds, falling from 8.4% to 3.1% after 1000 rounds. This visual representation reveals how AdaBoost significantly improves generalization compared to a single classifier, surpassing bagging in maintaining low test error even after achieving perfect training accuracy. The graphical results provide an intuitive guide for the reader to grasp how boosting and bagging maintain high model accuracy without overfitting.

Furthermore, the image below provides an example of a typical figures provided in the paper. We highlight the figure's simplicity and exhaustive description, showcasing it's understandability.

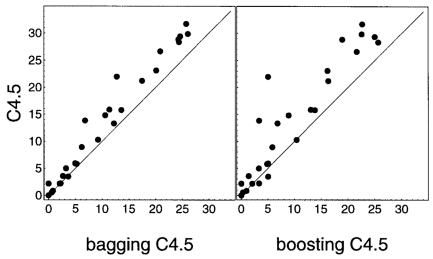


FIG. 2. *Comparison of* C4.5 *versus bagging* C4.5 *and boosting* C4.5 *on a set of* 27 *benchmark problems as reported in* [18]. *Each point in each scatter plot shows the test error rate of the two competing algorithms on a single benchmark. The y-coordinate of each point gives the test error rate* (*in percent*) *of* C4.5 *on the given benchmark, and the x-coordinate gives the error rate of bagging* (*left plot*) *or boosting* (*right plot*). *All error rates have been averaged over multiple runs.*

(1)

Applications and Universality

Understanding the empirical success of boosting algorithms is an important theoretical problem in machine learning. One of the most influential works presented in the paper is margin theory, which provides a series of upper, although loose, bounds for the generalization error of any voting classifier in terms of the margins of the training data. Margin theory is crucial in boosting algorithms, as it explains why these methods achieve strong generalization performance. This behavior of high generalization despite high complexity is explained through maximizing the margin of each

training example, which correlates with the ability of the classifier to generalize to unseen data. By focusing on increasing the margins, boosting algorithms like AdaBoost emphasize difficult-to-classify examples and refine their decision boundaries, reducing overfitting and improving robustness.

In practice, boosting and margin theory have found wide applications in various real-world domains, from image classification to fraud detection. The combination of theoretical upper bounds and empirical results has made boosting a preferred method in many machine learning tasks. Research has evolved significantly since Schapire *et al.*'s influential paper "Boosting the Margin," with numerous studies expanding on margin theory and its implications for boosting algorithms. Despite the progress, the fundamental principles laid out in that paper still hold relevance today. With ever more sophisticated machine learning techniques being created, the concepts of margin maximization and improving generalization continue to form the bedrock of many advanced machine learning algorithms used in real world scenarios.

3 Critic Review

Empirical tests play a crucial role in substantiating the theoretical claims made in the paper. Schapire *et al* use a variety of datasets to demonstrate that boosting not only improves error rates but also enhances margin distributions compared to other ensemble methods like bagging. This is evident in their experiments with the 'letter' dataset, where they show that "boosting drives the test error down even further to just 3.1 percent" (Schapire et al), a substantial improvement over other methods. The results presented are robust and the graphical representations of margin distributions provide a clear visual of how boosting affects classifier behavior over iterations. While the empirical evidence presented in the paper supports the margin theory, the experiments are somewhat limited in scope and diversity. The datasets used, such as the 'letter' dataset, do not necessarily represent the complexity or the types of data challenges present in current machine learning tasks. Modern datasets in areas like image recognition, natural language processing, or genomic classification might yield different results under the boosting framework proposed in the paper.

Weak Points and Simplifying Assumptions

The paper relies on assumptions that may not always hold true in real-world scenarios. The theoretical framework assumes an idealized behavior of classifiers and data distributions that may not be applicable to practical applications. As the authors acknowledge, "our bounds are too loose to give practical quantitative predictions," revealing limitations in their theoretical model. Furthermore, the paper does not explore situations where boosting might falter or margin assumptions may break down. A discussion on these limitations would provide a more balanced view of the applicability of their theory.

Scalability and Limitations of Synthetic Data

The paper discusses scalability but lacks comprehensive empirical data on how boosting performs in varied, large-scale environments. The paper does not provide extensive empirical data on the performance of boosting in varied or extremely large-scale environments. Addressing these gaps could enhance the paper's relevance and applicability in current machine-learning challenges. As datasets grow in size and complexity, the ability of learning algorithms to not only handle this scale but also maintain or improve performance is crucial. The margin-based framework proposed by Schapire et al. suggests that boosting can indeed scale effectively by focusing on increasing the margin, which theoretically should improve generalization error across larger datasets. However, the practical implications of this scalability are not fully explored in the paper. For example, the computational cost of scaling boosting methods, particularly in terms of processing time and memory requirements, is not adequately addressed. This is a critical oversight, as the efficiency of an algorithm is as important as its effectiveness, especially in large-scale applications.

Moreover, the use of synthetic data in experiments is primarily aimed at validating the theoretical claims regarding the effectiveness of boosting algorithms through controlled experiments. The paper explains, "In order to analyze the generalization error, one should consider more than just the training error... One should also take into account the confidence of the classifications" (Schapire *et al.* [1]). By using synthetic data, the authors can create scenarios that precisely test this hypothesis. While the use of synthetic data in machine learning research is a common practice that offers several advantages, it also introduces certain limitations, particularly in terms of the generalizability of the results. Synthetic data, often does not capture the same level of noise, outliers, and feature dependencies found in natural datasets. This limitation is critical in the context of boosting algorithms, which are known for their ability to handle complex and high-dimensional data. The synthetic data used might not fully challenge the boosting algorithms in ways that real-world data would, potentially leading to overly optimistic conclusions about their effectiveness. The parameters of boosting, such as the number of iterations and the learning rate, might be tuned to perform optimally on synthetic data but may not perform as well on real-world data. This discrepancy raises concerns about the practical applicability of the findings.

4 Additional Statements and Improvements

Throughtout the paper, Breiman and Fraud [1] prove theoretical bounds for the generalization of boosting based tree methods. However, to their own admission, these bounds are not tight and are often inexact as it pertains to practical implementations. There most notable conclusion was given by the following:

Theorem 1. *Let $\mathscr{D}$ be a distribution over $X \times \{-1, 1\}$, and let S be a sample of m examples chosen independently at random according to $\mathscr{D}$. Assume that the base-classifier space $\mathscr{H}$ is finite, and let $\delta > 0$. Then with probability at least $1 - \delta$ over the random choice of the training set S, every weighted average function $f \in \mathscr{C}$ satisfies the following bound for all $\theta > 0$:*

$$\mathbb{P}_{\mathscr{D}}[yf(x) \leq 0] \leq \mathbb{P}_S[yf(x) \leq \theta] + \mathcal{O}\left(\frac{1}{\sqrt{m}}\left(\frac{\log m \cdot \log |\mathscr{H}|}{\theta^2} + \log(1/\delta)\right)^{1/2}\right).$$

Advancement of Theorems

We provide an extension of this theorem by showing that the probability of the difference between any two weighted average classification models decreases exponentially as the number of classifiers N increases:

Theorem 2. *For $f \in \mathscr{C}(\mathscr{H})$ and $g \in \mathscr{C}_N(\mathscr{H})$ drawn i.i.d. according to distribution $\mathscr{Q}(f)$, we have*

$$\mathbb{P}_{S,g\sim\mathscr{Q}(f)}\left[yg(x) - yf(x) \geq \gamma\right] \leq \exp\left(\frac{-N\gamma^2}{2 - 2E_S^2[yf(x)] + 4\gamma/3}\right).$$

We first provide the following useful lemmas.

Lemma 1 (Markov's Inequality)**.** *Let X be a non-negative random variable with finite expected value $\mathbb{E}[X]$. Then, for any $\alpha > 0$,*

$$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Lemma 2 (Jensen's Inequality)**.** *Let X be a random variable, and let ϕ be a convex function in which the expectation of $\phi(X)$ is finite. Then,*

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

Proof. Take $\alpha > 0 \in \mathbb{R}^+$. By Lemma[1] with an exponential function, we have:

$$\mathbb{P}_{S,g\sim\mathscr{Q}(f)}\left[yg(x) - yf(x) \geq \gamma\right] \leq \exp(-\alpha N\gamma/2)\mathbb{E}_{S,g\sim\mathscr{Q}(f)}\left[\exp(\alpha(yg(x) - yf(x))/2)\right].$$

Since $g(x)$ is an average of N classifiers defined over a convex hull (given by Shepiere et al.), and the classifiers are drawn i.i.d., we can factorize the expectation and by applying Lemma 2 & independence of the random variables we now have:

$$\mathbb{E}_{S,g\sim\mathscr{Q}(f)}\left[\exp(\alpha(yg(x) - yf(x))/2)\right] = \prod_{i=1}^{N}\mathbb{E}_{S,h_i\sim\mathscr{Q}(f)}\left[\exp(\alpha(yh_i(x) - yf(x))/2)\right].$$

Each expectation term can be approximated using a Taylor series expansion:

$$\mathbb{E}_{S,h_j\sim\mathscr{Q}(f)}\left[\exp(\alpha(yh_j(x) - yf(x))/2)\right] \approx 1 + \frac{\alpha^2}{6}\mathbb{E}_{S,h_j\sim\mathscr{Q}(f)}\left[(yh_j(x) - yf(x))^2\right].$$

Given that the classifiers are bounded in the range $\{-1, +1\}$, the variance of their predictions is also bounded. Thus, we have:

$$\mathbb{E}_{S,h_j\sim\mathscr{Q}(f)}\left[(yh_j(x) - yf(x))^2\right] = 1 - E_S^2[yf(x)],$$

where $E_S^2[yf(x)]$ represents the squared expected value of the classifier output.

Substituting this back into the Taylor expansion approximation, we obtain:

$$\mathbb{E}_{S,h_j\sim\mathscr{Q}(f)}\left[\exp(\alpha(yh_j(x) - yf(x))/2)\right] \approx \exp\left(\frac{\alpha^2}{6}(1 - E_S^2[yf(x)])\right).$$

Thus, the overall expectation simplifies to:

$$\mathbb{E}_{S,g\sim\mathscr{Q}(f)}\left[\exp(\alpha(yg(x)-yf(x))/2)\right] \approx \exp\left(\frac{\alpha^2 N}{6}(1-E_S^2[yf(x)])\right).$$

Substituting back into our original inequality, we have:

$$\mathbb{P}_{S,g\sim\mathscr{Q}(f)}\left[yg(x)-yf(x)\geq\gamma\right] \leq \exp\left(-\alpha N\gamma/2+\frac{\alpha^2 N}{6}(1-E_S^2[yf(x)])\right).$$

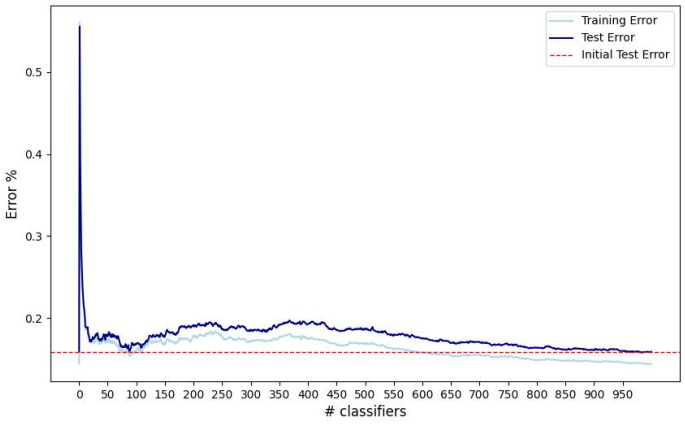
Choosing $\alpha = \frac{3\gamma}{2-2E_S^2[yf(x)]+4\gamma/3}$, we obtain the desired result:

$$\mathbb{P}_{S,g\sim\mathscr{Q}(f)}\left[yg(x)-yf(x)\geq\gamma\right] \leq \exp\left(\frac{-N\gamma^2}{2-2E_S^2[yf(x)]+4\gamma/3}\right).$$

Therefore we conclude that the probability of the difference between predictions of two functions, g and f, exceeding a threshold γ, is exponentially bounded. □

Evaluation on Present Dataset

Given the outdated methodologies of Schapire et al., we reassess the validity of boosting-based generalization in complex models using the MNIST handwriting dataset, a more difficult popular classification problem.



(2)

The graph illustrates how AdaBoost's error rates change with increasing classifiers on the MNIST dataset. Training error stabilizes early at a low level, while test error gradually declines, confirming the findings in "Boosting the Margin." Adding more classifiers strengthens margins and robustness, enhancing predictive accuracy while reducing overfitting, which underscores the effectiveness of margin theory.

5 Conclusion and Future Work

Throughout the paper, Schapire and his co-authors established prove theoretical bounds for the generalization of boosting based tree methods which, to their own admission, are not practically tight and are often inexact as it pertains to practical implementations. However, given the ingenious idea of margin theory introduced and theorems explained by this paper, we highlight its relevance not only to the margin theory field but also to algorithmic development in machine learning disciplines, specifically for multi-classification problems. As more models are trained and the understanding of boosting improves, this work continues to be pivotal, underlining the importance of robust ensemble methods in enhancing the predictive performance of machine learning algorithms. We advocate for the investigation of more tighter generalization bounds and how further analysis of how generalization evolves under non-finite classification spaces.

References

[1] Schapire, Robert E., et al. “Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods.” The Annals of Statistics, vol. 26, no. 5, 1998, pp. 1651–86. JSTOR, http://www.jstor.org/stable/120016. Accessed May 11, 2024.

[2] Antos, A., K´egl, B., Linder, T., Lugosi, G., 2002. Data-dependent marginbased generalization bounds for classification. Journal of Machine Learning Research 3, 73–98

[3] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, vol. 86, no. 11, 1998, pp. 2278-2324.

[4] Vapnik, V. N., 1998. Statistical Learning Theory. John Wiley & Sons, New York