Artificial Intelligence 203 (2013) 1–18

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

On the doubt about margin explanation of boosting

Wei Gao, Zhi-Hua Zhou *

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Article history:
Received 8 May 2012
Received in revised form 4 July 2013
Accepted 13 July 2013
Available online 24 July 2013

Keywords:
Classification
Boosting
Ensemble methods
Margin theory

ABSTRACT

Margin theory provides one of the most popular explanations to the success of AdaBoost, where the central point lies in the recognition that *margin* is the key for characterizing the performance of AdaBoost. This theory has been very influential, e.g., it has been used to argue that AdaBoost usually does not overfit since it tends to enlarge the margin even after the training error reaches zero. Previously the *minimum margin bound* was established for AdaBoost, however, Breiman (1999) [9] pointed out that maximizing the minimum margin does not necessarily lead to a better generalization. Later, Reyzin and Schapire (2006) [37] emphasized that the margin distribution rather than minimum margin is crucial to the performance of AdaBoost. In this paper, we first present the *kth margin bound* and further study on its relationship to previous work such as the minimum margin bound and Emargin bound. Then, we improve the previous empirical Bernstein bounds (Audibert et al. 2009; Maurer and Pontil, 2009) [2,30], and based on such findings, we defend the margin-based explanation against Breiman's doubts by proving a new generalization error bound that considers exactly the same factors as Schapire et al. (1998) [39] but is sharper than Breiman's (1999) [9] minimum margin bound. By incorporating factors such as average margin and variance, we present a generalization error bound that is heavily related to the whole margin distribution. We also provide margin distribution bounds for generalization error of voting classifiers in finite VC-dimension space.

1. Introduction

The AdaBoost algorithm [18,19], which aims to construct a "strong" classifier by combining some "weak" learners (slightly better than random guess), is a representative of ensemble methods [47] and has been one of the most influential classification algorithms [13,45], and it has exhibited excellent performance both on benchmark datasets and real applications [5,16].

Many studies are devoted to understanding the mysteries behind the success of AdaBoost, among which the margin theory proposed by Schapire et al. [39] has been very influential. For example, AdaBoost often tends to be empirically resistant (but not completely) to overfitting [8,17,35], i.e., the generalization error of the combined learner keeps decreasing as its size becomes very large and even after the training error has reached zero; it seems violating the Occam's razor [7], i.e., the principle that less complex classifiers should perform better. This remains one of the most famous mysteries of AdaBoost. The margin theory provides the most intuitive and popular explanation to this mystery, that is: AdaBoost tends to improve the margin even after the error on training sample reaches zero.

However, Breiman [9] raised serious doubt on the margin theory by designing arc-gv, a boosting-style algorithm. This algorithm is able to maximize the *minimum margin*, i.e., the smallest margin over the training data (the formal definition

* Corresponding author.
E-mail address: zhouzh@lamda.nju.edu.cn (Z.-H. Zhou).

0004-3702/$ – see front matter
http://dx.doi.org/10.1016/j.artint.2013.07.002

will be given in Eq. (2)), but its generalization error is high on empirical datasets, and similar experimental evidence has also been observed in [22]. Thus, Breiman [9] concluded that the margin theory for AdaBoost failed. Breiman's argument was backed up with a minimum margin bound, which is sharper than the generalization bound given by Schapire et al. [39], and a lot of experiments. Garg and Roth [21] presented a margin-distribution algorithm based on a data-dependent complexity measure. Later, Reyzin and Schapire [37] found that there were flaws in the design of experiments: Breiman used CART trees [11] as base learners and fixed the number of leaves for controlling the complexity of base learners. However, Reyzin and Schapire [37] found that the trees produced by arc-gv were usually much deeper than those produced by AdaBoost. Generally, for two trees with the same number of leaves, the deeper one is with a larger complexity because more judgments are needed for making a prediction. Therefore, Reyzin and Schapire [37] concluded that Breiman's observation was biased due to the poor control of model complexity. They repeated the experiments by using decision stumps for base learners, considering that decision stump has exactly two leaves and thus with a fixed complexity, and observed that though arc-gv produced a larger minimum margin, its margin distribution was quite poor. Nowadays, it is well-accepted that the margin distribution is crucial to relate margin to the generalization performance of AdaBoost. To support the margin theory, Wang et al. [44] presented a sharper bound in term of *Emargin* (the formal definition will be given in Theorem 3), which was believed to be relevant to margin distribution.

In this paper, we first present the *kth margin* bound and further study its relationship to previous work such as the minimum margin bound and Emargin bound. Then, by using empirical Bernstein bounds, we present a new generalization error bound for voting classifier, which considers exactly the same factors as Schapire et al. [39], but is sharper than the bounds of Schapire et al. [39] and Breiman [9]. Therefore, we defend the margin-based explanation against Breiman's doubt. Moreover, we provide a generalization error bound, by incorporating other factors such as average margin and variance, which are heavily relevant to the whole margin distribution. We also give a margin distribution bound for generalization error of voting classifiers in finite VC-dimension space. It is also worth mentioning that our new empirical Bernstein bounds improve the main results of [2,30], with a simpler proof, and we present empirical Bernstein bounds for finite VC-dimension space; these results can be interesting, independently to the main purpose of the paper, to the machine learning community.

The rest of this paper is organized as follows. We begin with some notations and background in Sections 2 and 3, respectively. Then, we prove the kth margin bound and discuss on its relation to previous bounds in Section 4. Our main results are presented in Section 5, and detailed proofs are provided in Section 6. We conclude in Section 7.

2. Notations

Let $\mathcal{X}$ and $\mathcal{Y}$ denote an input space and output space, respectively. In this paper, we focus on binary classification problems, i.e., $\mathcal{Y} = \{+1, -1\}$. Denote by D an (unknown) underlying probability distribution over the product space $\mathcal{X} \times \mathcal{Y}$. A training sample of size m

$$S = \big\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\big\}$$

is drawn independently and identically (i.i.d.) according to the distribution D. We use $\Pr_D[\cdot]$ to refer as the probability with respect to D, and $\Pr_S[\cdot]$ to denote the probability with respect to uniform distribution over the sample S. Similarly, we use $E_D[\cdot]$ and $E_S[\cdot]$ to denote the expected values, respectively. For an integer $m > 0$, we set $[m] = \{1, 2, \ldots, m\}$.

The Bernoulli Kullback–Leibler (or KL) divergence is defined as

$$KL(q\|p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \quad \text{for } 0 \leqslant p, q \leqslant 1.$$

For a fixed q, we can easily find that $KL(q\|p)$ is a monotone increasing function for $q \leqslant p < 1$, and thus, the inverse of $KL(q\|p)$ for the fixed q is given by

$$KL^{-1}(q; u) = \inf_w \big\{w\colon\ w \geqslant q \text{ and } KL(q\|w) \geqslant u\big\}.$$

Let $\mathcal{H}$ be a hypothesis space. A base learner is a function which maps a distribution over $\mathcal{X} \times \mathcal{Y}$ onto a function $h : \mathcal{X} \to \mathcal{Y}$. In this paper, we only focus on binary base classifiers, i.e., the outputs are in $\{-1, 1\}$. Let $\mathcal{C}(\mathcal{H})$ denote the convex hull of H, i.e., a voting classifier $f \in \mathcal{C}(\mathcal{H})$ is of the following form

$$f = \sum \alpha_i h_i \quad \text{with } \sum \alpha_i = 1 \text{ and } \alpha_i \geqslant 0.$$

For $N \geqslant 1$, denote by $\mathcal{C}_N(\mathcal{H})$ the set of unweighted averages over N elements from $\mathcal{H}$, that is

$$\mathcal{C}_N(\mathcal{H}) = \left\{g\colon\ g = \sum_{j=1}^{N} \frac{h_j}{N},\ h_j \in \mathcal{H}\right\}. \tag{1}$$

For voting classifier $f \in \mathcal{C}(\mathcal{H})$, we can associate with a distribution over $\mathcal{H}$ by using the coefficients $\{\alpha_i\}$, denoted by $\mathcal{Q}(f)$. For convenience, $g \in \mathcal{C}_N(\mathcal{H}) \sim \mathcal{Q}(f)$ implies $g = \sum_{j=1}^{N} h_j/N$ where $h_j \sim \mathcal{Q}(f)$.