Journal of Computer Science and Cybernetics, V.28, N.4 (2012), 310–322

BOUNDING AN ASYMMETRIC ERROR OF A CONVEX COMBINATION OF CLASSIFIERS

PHAM MINH TRI[1], CHAM TAT JEN[2]

[1]*Cambridge Research Laboratory, Toshiba Research Europe Ltd, Cambridge, United Kingdom; Email: mtpham@crl.toshiba.co.uk*
[2]*School of Computer Engineering, Nanyang Technological University, Singapore*

Tóm tắt. Sai số phân loại bất đối xứng là loại sai số trong đó có sự thỏa hiệp giữa tỷ lệ dương tính giả và tỷ lệ âm tính giả của bộ phân loại nhị phân. Nó được sử dụng rộng rãi gần đây nhằm giải quyết bài toán phân loại nhị phân mất cân đối, ví dụ phương pháp thúc đẩy bất đối xứng (asymmetric boosting) trong máy học. Tuy nhiên, cho đến nay, mối quan hệ giữa sai số bất đối xứng thực nghiệm và sai số bất đối xứng tổng quát chưa được giải quyết triệt để. Các cận cổ điển của sai số phân loại thông thường (sai số đối xứng) không dễ được áp dụng trong trường hợp mất cân đối, vì tỷ lệ dương tính giả và tỷ lệ âm tính giả được gán những chi phí khác nhau, và xác suất mỗi loại không được phản ánh bởi tập dữ liệu tập huấn. Trong bài báo này, chúng tôi trình bày một dạng cận trên cho sai số bất đối xứng tổng quát dựa trên sai số bất đối xứng thực nghiệm, của bộ phân loại có dạng là kết hợp lồi của nhiều bộ phân loại khác. Bộ phân loại kết hợp lồi được sử dụng khá phổ biến trong các phương pháp kết hợp phân loại gần đây như phương pháp thúc đẩy (boosting) hoặc phương pháp đóng bao (bagging). Chúng tôi cũng chỉ ra loại cận này là một dạng tổng quát của một trong những cận mới nhất (và chặt nhất) của sai số đối xứng tổng quát, cho bộ phân loại kết hợp lồi.

Abstract. Asymmetric error is an error that trades off between the false positive rate and the false negative rate of a binary classifier. It has been recently used in solving the imbalanced classification problem, *e.g.,* in asymmetric boosting. However, to date, the relationship between an empirical asymmetric error and its generalization counterpart has not been addressed. Bounds on the classical generalization error are not directly applicable since different penalties are associated with the false positive rate and the false negative rate respectively, and the class probability is typically ignored in the training set. In this paper, we present a bound on the expected asymmetric error of any convex combination of classifiers based on its empirical asymmetric error. We also show that the bound is a generalization of one of the latest (and tightest) bounds on the classification error of the combined classifier.

Keywords. Asymmetric error, asymmetric boosting, imbalanced classification, Rademacher complexity

1. INTRODUCTION

In recent years, the imbalanced binary classification problem has received considerable attention in various areas such as machine learning and pattern recognition. A two-class data set is said to be imbalanced (or skewed) when one of the classes (the minority/positive one)

is heavily under-represented in comparison with the other class (the majority/negative one). This issue is particularly important in real-world applications where it is costly to mis-classify examples from the minority class. Examples include: diagnosis of rare diseases, detection of fraudulent telephone calls, face detection and recognition, text categorization, information retrieval and filtering tasks, examples and absence of rare cases, respectively.

The traditional classification error is typically not used to learn a classifier in an imbalanced classification problem. In many cases, the probability of the positive class is a very small number. For instance, the probability of a face sub-window in appearance-based face detection (*e.g.,* Viola–Jones [28]) is less than 10^{-6}, while the probability of a non-face sub-window is almost 1. Using the classification error for learning would result in a classifier that has a very low false positive rate and a near-one false negative rate.

A number of cost-sensitive learning methods have been proposed recently to learn an imbalanced classifier. Instead of treating the given (labelled) examples equally, these methods introduce different weights to the examples of different classes of the input data set, so that one type of error rate can be reduced at the cost of an increase in the other type. These methods have appeared in a number of popular classification learning techniques, including: decision trees [1, 9], neural networks [14], support vector machines [26], and boosting [8, 15, 29].

Because the positive class is much smaller than the negative class, it is expensive to maintain a very large set of negative examples together with a small set of positive examples so as to have *i.i.d.* training examples. In practice, we typically have a fixed-size set of *i.i.d.* training examples for each class instead. In other words, *the class probability is ignored.* Let $f : \mathcal{X} \to \mathbb{R}$ represent the classifier with which we use $sign(f(x)) \in \mathcal{Y} = \{-1, +1\}$ to predict the class of x. By incorporating the weights into the learning process, these methods learn a classifier by minimizing the following *asymmetric error*,

$$\lambda_1 \mathbb{P}\left(f(x) \leq 0 | y = 1\right) + \lambda_2 \mathbb{P}\left(f(x) \geq 0 | y = -1\right), \tag{1.1}$$

where $\lambda_1, \lambda_2 > 0$ are the associated costs for each error rate: the false positive rate $\mathbb{P}\left(f(x) \leq 0 | y = 1\right)$ and the false negative rate $\mathbb{P}\left(f(x) \geq 0 | y = -1\right)$, and $\mathbb{P}$ is a probability measure on $\mathcal{X} \times \mathcal{Y}$ that describes the underlying distribution of instances and their labels. Note that the asymmetric error is a generalization of the classification error. One can obtain the classification error by choosing $\lambda_1 = \mathbb{P}(y = 1)$ and $\lambda_2 = \mathbb{P}(y = -1)$.

The motivation of the presented work comes from the success of recent real-time face detection methods in computer vision. These methods follow a framework proposed by Viola and Jones [28], in which a cascade of coarse-to-fine convex combinations of weak classifiers (or combined classifiers for short) is learned. At first, the combined classifiers were learned using AdaBoost [4]. However, recent advances [29, 15, 19] show that the accuracy and the speed of face detection could be significantly improved by replacing AdaBoost with asymmetric boosting [29], a variant of AdaBoost adapted to the imbalanced classification problem by minimizing 1.1.

Our work is inspired by the work in [19]. In this work, the authors showed that by choosing asymmetric costs λ_1, λ_2 such that $\frac{\lambda_1}{\lambda_2} = \frac{\alpha}{\beta}$, asymmetric boosting can obtain a classifier such that its false positive rate is less than α, its false negative rate is less than β, and *the number of weak classifiers is approximately minimized.* The first two results are necessary for the construction of a cascade. However, the third result is crucial because in real-time object detection, the number of weak classifiers is inversely proportional to the detection speed.

The success of real-time face detection has attracted a lot of attention as of late. However, there has been no theoretical explanation on the performance of asymmetric boosting. It is important to answer this question because there are new machine learning methods that rely on the knowledge about the generalization of the classifier to operate and improve it over time. Examples are online learning (*e.g.,* [7, 18]) and semi-supervised boosting (*e.g.,* [5]) for object detection. Existing bounds on the classification error cannot be applied here, because in this context the input data are treated as *per-class i.i.d.* examples, and we have different costs associated with the two classes. The goal for this work is, therefore, to develop bounds on (1.1) with respect to empirical errors to explain the performance of a combined classifier learned in the imbalanced case, *e.g.,* by using asymmetric boosting.

The outline of the paper is as follows. Section 2 gives a brief review of related work. The main results are presented in section 3. Conclusions are given in section 4. The proofs for the main results are given in section 5.

2. RELATED WORK

Let us focus our attention on work related to bounding the expected classification error of a combined classifier. Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a set of n training examples, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Under the *i.i.d.* assumption on the training examples, the standard approach to bounding the classification error was developed in seminal papers of Vapnik and Chervonenkis in the 70s and 80s (*e.g.,* see [3, 25, 27]). The bounds are expressed in terms of the empirical probability measure and the VC-dimension of the function class. However, in many important examples (*e.g.,* in boosting or in neural network learning), directly applying these bounds would not be too useful because the VC-dimension of the function class can be very large, or even infinite.

Since the invention of voting algorithms such as boosting, the convex hull,

$$conv\,(\mathcal{H}) := \left\{ \sum_{i=1}^{\infty} w_i h_i : w_i \geq 0, \sum_{i=1}^{\infty} w_i = 1, h_i \in \mathcal{H} \right\}, \tag{2.2}$$

of a base function class $\mathcal{H} := \{h : \mathcal{X} \to [-1, 1]\}$ has become an important object of study in the machine learning literature. This is because: (1) $conv\,(\mathcal{H})$ represents the space of all linear ensembles of base functions in $\mathcal{H}$, and (2) traditional techniques using VC-dimension cannot be applied directly because even if the base class $\mathcal{H}$ has a finite VC-dimension, the combined class $\mathcal{F}$ has an infinite VC-dimension.

Schapire *etal.* [20, 21] pioneered a line of research to explain the effectiveness of voting algorithms. They developed a new class of bounds on the classification error of a convex combination of classifiers, expressed in terms of the empirical distribution of margins $yf(x)$. They showed that in many experiments, voting methods tend to classify examples with large margins.

Koltchinskii *etal.* [12, 13, 10] combined the theories of empirical, Gaussian, and Rademacher processes to refine this type of bounds. They used Talagrand's remarkable inequalities on empirical processes, exploiting subsets of the convex hull to which the classifier belongs, the sparsity of the weights, and the clustering properties of the weak classifiers, to further tighten the bounds. Some of these properties are related to the learning algorithm that was used to learn the combined classifier.

To the best of our knowledge, little work related to bounding the expected asymmetric error defined in (1.1) has been done. In [2, 22], the authors targeted at bounding the Neyman-Pearson error of a classifier, with respect to the VC-dimension of the function class. The Neyman-Pearson error is fundamentally different from (1.1). Consider two error rates: the false positive rate and the false negative rate. In the former, one constrains one error rate and minimizes the other; in the latter, one minimizes a weighted sum of the two error rates. Besides, [2, 22] are not suitable to explain a combined classifier learned from boosting, because in this case, the classifier's VC-dimension is possibly infinite.

Zadrozny *etal.* [30] proposed a method to convert a classification learning algorithm into a cost-sensitive one, and proved that the resultant cost-sensitive error is at most M times the resultant classification error were the classifier learned with the original algorithm, where M is approximately inversely proportional to the probability of the positive class. One can apply their work on the bounds of Koltchinskii *etal.* to obtain bounds on (1.1). However, factor M is too large in practice because the probability of the positive class is too small. For instance, in the context of face detection we are interested in, $M \approx 10^6$, implying the resultant bound is loosened by 10^6 times.

3. MAIN RESULTS

In this paper, we propose bounds which are generalizations of Theorem 1 and Corollary 1 of Koltchinskii and Panchenko [12]. Theorem 1 of [12] is one of the tightest bounds to date on the classification error of a combined classifier. However, they cannot be trivially generalized because they operate on the assumption that the training examples are *i.i.d.* and are treated equally. At the centre of the study presented in [12] is the result of Panchenko [16] on the deviation of an empirical process. We propose a new result that is the generalization of [16]'s work. It allows to include weights on the examples, and to eliminate the *identical* requirement on the training set. By using the new result, we are able to generalize the work of [12] to bound the expected asymmetric error of a combined classifier.

There are tighter bounds in [12] which operate under more restricted assumptions on the combined classifier. However, studying them is beyond the scope of this paper. We leave that for future work.

In our method, we do not need to convert a learning algorithm, avoiding the problem of loosening the bound by M times as in [30].

The contribution of the paper can be summarized as follows. In [12], Koltchinskii and Panchenko derived their generalization bounds based on Panchenko's study [16] on the deviation of an empirical process. We generalize [16] by introducing weights on the examples, so that we can incorporate different costs to different classes. We then specialize the result in the context of bounding an asymmetric error, using the strategy that [16] was specialized to derive the bounds in [12]. Most of our derivations are minor variations on some proofs in [17, 16, 12]. Our only claim of originality is for the recognition that an expected asymmetric error can be bounded by its empirical asymmetric error in the same way that the expected classification error is bounded by its empirical error.

Suppose that $\mathbb{P}$ is a probability measure on $\mathcal{X} \times \mathcal{Y}$, which describes the underlying distribution of instances and their labels. Let $\mathbb{P}_v$ be the probability measure on some random variable v given that other random variables are fixed. We denote by $\mathbb{E}$ and $\mathbb{E}_v$ their expectations, re-

spectively. Suppose that the training set x consists of n_1 positive examples $\{x_1, \ldots, x_{n_1}\}$ and n_2 negative examples $\{x_{n_1+1}, \ldots, x_n\}$ where $n = n_1 + n_2$. Let $\mathcal{F} := \{f : \mathcal{X} \to [-q/2, q/2]\}$ be a function class for some $q > 0$. Panchenko [16] studied the deviation of a functional

$$p_n f := \frac{1}{n} \sum_{i=1}^{n} f(x_i), \tag{3.3}$$

from its mean $\mathbb{E}[p_n f]$, under the standard assumption that the variables x_i for $i = 1..n$ are drawn identically and independently from a probability measure μ on $\mathcal{X}$. In our case, we consider the deviation of a more general functional,

$$P_n f := \sum_{i=1}^{n} a_i f(x_i), \tag{3.4}$$

from its mean $Pf := \mathbb{E}[P_n f]$, for some known $a_i \in \mathbb{R}$ for all $i = 1..n$, and under an assumption that x_i are drawn independently, but not necessarily identically. The weights a_i allow us to associate different costs to different examples, a general condition often needed in the imbalanced classification context. The elimination of the identical condition is required, since in the imbalanced case, positive examples and negative examples are typically not drawn from the same distribution (the class probability is ignored).

However, this requirement does not really pose a difficulty because most standard techniques in bounding empirical processes do not require the identical condition.

We control the residual $Q_n f := Pf - P_n f$ uniformly over the function class $\mathcal{F}$ by using the same measure propopsed in [16] called *uniform packing number*. We need some definitions. Let $W_n f(y) := \sum_{i=1}^{n} a_i^2 (f(y_i) - f(x_i))^2$ be a function that measures how the given training set x differs from another training set y (of n examples) under the action of f. Let $V_n f := \mathbb{E}_y W_n f(y)$ be its expectation over all y. Given a probability distribution Q on $[-q/2, q/2]$, let us denote by $d_{Q,2}(f, g) := (Q(f - g)^2)^{1/2}$ the $L_2(Q)$-distance in $\mathcal{F}$. Given $u > 0$, a subset $\mathcal{F}' \subseteq \mathcal{F}$ is called u-separated if for any pair $f \neq g \in \mathcal{F}'$, we have $d_{Q,2}(f, g) > u$. Let the *packing number* $D(\mathcal{F}, u, d_{Q,2})$ be the maximal cardinality of any u-separated set. Let the *uniform packing number* $D(\mathcal{F}, u)$ be a function such that $\sup_Q D(\mathcal{F}, u, d_{Q,2}) \leq D(\mathcal{F}, u)$ where the supremum is taken over all probability measures on $\mathcal{X}$. We say that $\mathcal{F}$ satisfies the uniform entropy condition if

$$\int_0^\infty \sqrt{\log D(\mathcal{F}, u)} du < \infty. \tag{3.5}$$

Our first new result, stated in Theorem 3.1, is a generalization of Corollary 3 presented in [16]. The proof of Theorem 3.1 is given in section 5.1.

Theorem 3.1. *If* (3.5) *holds, for any training set x of n examples and any $\beta \in (0, 1)$, there exists a constant $0 < K < \infty$ that depends on β only such that for any $t \geq \log \beta^{-1}$, with probability at most* $\exp\left(1 - (\sqrt{t} - \sqrt{\log \beta^{-1}})^2\right)$,

$$\exists f \in \mathcal{F}, Q_n f \geq K \int_0^{\frac{\sqrt{V_n f}}{2m}} \sqrt{\log D(\mathcal{F}, u)} du + \sqrt{4 V_n f t}, \tag{3.6}$$

where $m := \sqrt{\sum_{i=1}^{n} a_i^2}$.

By using Theorem 3.1, we derive our second result, stated in Theorem 3.2. Theorem 3.2 serves as an intermediate result so that we can derive bounds on the asymmetric error in the final step. It is also a generalization of Theorem 1 in [12]. From now on, we assume $a_i > 0$ for all $i = 1..n$ without loss of generality (since if $a_i < 0$, we can replace h_i with $-h_i$). As with Theorem 1 in [12], the bound is controlled by the γ-dimension introduced in [10], described as follows. Assume that the weights of the weak classifiers are arranged in the decreasing order, $w_1 \geq w_2 \geq \ldots$. For a number $\gamma \in [0, 1]$, a *γ-dimension* of f, denoted as $d(f; \gamma)$, is defined as the smallest integer $d \geq 0$ such that there exists $T \geq 0$, weak classifiers $h_j \in \mathcal{H}$, and weights $w_j \geq 0$ such that $f = \sum_{j=1}^{T} w_j h_j$, $\sum_{j=1}^{T} w_j \leq 1$, and $\sum_{j=d+1}^{T} w_j \leq \gamma$. Given a family of weak classifiers $\mathcal{H}$, we assume, for some $V > 0$,

$$D(\mathcal{H}, u) = O(u^{-V}). \tag{3.7}$$

This is often the case in practice. For instance, if $\mathcal{H}$ is a VC-subgraph class with VC-dimension d, then by the well-known result of Dudley and Pollard (*e.g.,* [6]), (3.7) holds with $V = 2d$, namely $D(\mathcal{H}, u) \leq e(d + 1)(2e/u^2)^d$.

Let $\varphi_\delta : \mathbb{R} \to [0, 1]$ be a Lipschitz function with Lipschitz constant δ^{-1} (*ie.* $\frac{|\varphi_\delta(s_1) - \varphi_\delta(s_2)}{s_1 - s_2 | \leq \delta^{-1} \forall s_1, s_2 \in \mathbb{R}}$). For some function $f \in \mathcal{F}$, let us define a function $f_{\varphi_\delta}(x, y) := \varphi_\delta(yf(x))$. We present the following theorem.

Theorem 3.2. *Let $\alpha = 2V/(V + 2)$. If (3.7) holds, then for all $t > 0$, for all $f \in \mathcal{F} := conv(\mathcal{H})$, and for all $\varphi_\delta : \mathbb{R} \to [0, 1]$ as a Lipschitz function with Lipschitz constant δ^{-1} where $\delta \in \Delta = \{2^{-k} : k \geq 1\}$, with probability at least $1 - e^{-t}$, the following inequality holds*

$$\frac{Pf_{\varphi_\delta} - P_n f_{\varphi_\delta}}{\sqrt{Pf_{\varphi_\delta}}} \leq K \inf_{\gamma \in [0,1]} \left\{ \sqrt{\frac{d(f; \gamma)}{n'} \log \frac{n'}{\delta}} + \left(\frac{\gamma}{\delta}\right)^{\alpha/2} \frac{(Pf')^{-\alpha/4}}{\sqrt{n'}} + \sqrt{\frac{t}{n'}} \right\}, \tag{3.8}$$

where $n' = \min_{i=1..n}\{1/a_i\}$.

The proof for Theorem 3.2 is given in section 5.2. We now choose proper values for weights a_i to derive bounds on an expected asymmetric error with respect to its empirical asymmetric error. This is the final result of the paper. Let $a_i = \lambda_1/n_1$ for $1 \leq i \leq n_1$, and $a_i = \lambda_2/n_2$ for $n_1 + 1 \leq i \leq n$. Suppose for some function $g : \mathcal{X} \to \mathcal{Y}$, $\mathbb{P}_{n_1} g = \frac{1}{n_1} \sum_{i=1}^{n_1} g(x_i)$ and $\mathbb{P}_{n_2} g = \frac{1}{n_2} \sum_{i=n_1+1}^{n} g(x_i)$ denote the empirical probability measure of g on the positive set and the negative set, respectively. Our final result is the following corollary (see section 5.3 for the proof).

Corollary 3.1. *Let $\alpha = 2V/(V + 2)$ and $n' = \min\{n_1/\lambda_1, n_2/\lambda_2\}$. If (3.7) holds, then for all $t > 0$ with probability at least $1 - e^{-t}$ for all $f \in \mathcal{F} := conv(\mathcal{H})$, the following inequality holds,*

$$\begin{aligned} &\lambda_1 \mathbb{P}(f(x) \leq 0 | y = +1) + \lambda_2 \mathbb{P}(f(x) \geq 0 | y = -1) \\ &\quad \leq K \inf_{\delta \in (0,1]} \Bigg\{ \lambda_1 \mathbb{P}_{n_1}(f(x) \leq \delta) + \lambda_2 \mathbb{P}_{n_2}(f(x) \geq -\delta) \\ &\quad + \inf_{\gamma \in [0,1]} \left\{ \sqrt{\frac{d(f; \gamma)}{n'} \log \frac{n'}{\delta}} + \left(\frac{\gamma}{\delta}\right)^{\frac{2\alpha}{(2+\alpha)}} n'^{\frac{-2}{2+\alpha}} \right\} + \sqrt{\frac{t}{n'}} \Bigg\}. \end{aligned} \tag{3.9}$$

Corollary 3.1 bounds the expected asymmetric error $\lambda_1\mathbb{P}(f(x) \leq 0|y = +1) + \lambda_2\mathbb{P}(f(x) \geq 0|y = -1)$ with respect to its empirical margin-based asymmetric error $\lambda_1\mathbb{P}_{n_1}(f(x) \leq \delta) + \lambda_2\mathbb{P}_{n_2}(f(x) \geq -\delta)$. One can easily see that when $\lambda_1 = \mathbb{P}(y = +1)$ and $\lambda_2 = \mathbb{P}(y = -1)$, corollary 3.1 degenerates to Corollary 1 in [12].

Theorem 1 in [12] can be interpreted as interpolation between zero-error and nonzero-error cases of the classification error. Since theorem 3.2 is a generalization Theorem 1 in [12], it can be considered as interpolation between the zero case and the nonzero case of an asymmetric error.

To the best of our knowledge, the best bound on the expected asymmetric error, prior to our work, can only be obtained by applying Zadrozny's work [30] on top of a generalization bound like those in [11, 10, 12]. As discussed in section 2, this approach induces an extra large term M which is inversely proportional to $\mathbb{P}(y = +1)$. Our bound in Corollary 3.1 does not involve this term, while at the same time, it is a generalization (in the asymmetric context) of Corollary 1 in [12], one of the tightest bounds on the traditional classification error of a combined classifier. It is therefore reasonable to claim that our bound is M times tighter than the same kind of bound obtained from using the Zadrozny's approach [30], and that it is fairly tight overall.

4. CONCLUSION

In this paper, we have proposed a new set of bounds which contrain the expected asymmetric error of a combined classifier based on its margin-based empirical asymmetric error. Our bounds can be considered as generalizations (in the asymmetric context) of one of the tightest bounds on the classification error of a combined classifier, presented in [12]. It is shown that our bounds on the expected asymmetric error are tighter than previously best known bounds, derived from the approach introduced by [30]. In the future, we will focus on further tightening the proposed bounds, as well as developing bounds on the asymmetric error for the cases of online learning and semi-supervised learning a combined classifier (*e.g.,* [7, 18, 5]).

5. PROOFS

5.1. Proof for theorem 3.1

The proof is derived by combining the two following lemmas. Let us introduce the symmetrized version of Q_nf: $S_nf := \sum_{i=1}^n a_i(f(y_i) - f(x_i))$, and S_nf's randomized version: $R_nf = \sum_{i=1}^n \varepsilon_i a_i(f(y_i) - f(x_i))$, where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are random variables in $\mathcal{X}^n$, and $\varepsilon_1, \ldots, \varepsilon_n$ are *i.i.d.* Rademacher random variables (*ie.* $\varepsilon_i \in \{-1, 1\}$ with equal probabilities). For any pair $x, y \in \mathcal{X}^n$, let Π be the set of permutations π of these coordinates such that for each $i = 1..n$, $\pi(x_i), \pi(y_i) \in \{x_i, y_i\}$. Consider a function $\Phi_nf := \Phi(f, x, y)$ which is invariant to the permutations $\pi \in \Pi$ of (x, y). Assume that for some fixed $\beta \in (0, 1)$ and for any fixed $x, y \in \mathcal{X}^n$, we have

$$\mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}(R_nf - \Phi_nf) > 0\right) < 1 - \beta. \tag{5.10}$$

Then the following lemma holds.

Lemma 5.1. *If* (5.10) *holds, then for any* $t \geq \log \beta^{-1}$, *we have*

$$\mathbb{P}\left(\exists f \in \mathcal{F}, Q_n f \geq \mathbb{E}_y \Phi_n f + \sqrt{4V_n f t}\right) < \exp\left(1 - (\sqrt{t} - \sqrt{\log \beta^{-1}})^2\right). \tag{5.11}$$

Chứng minh. We will begin by stating some previous results needed for the proof.

Proposition 5.1. *[Corollary 1 of [16]]Let* $\xi_i(x, y) : \mathcal{X}^{2n} \to \mathbb{R}$, $1 \leq i \leq 3$, *be measurable functions defined on two copies of* $\mathcal{X}^n$ *and let* $\xi_i'(x) = \mathbb{E}_y[\xi_i(x, y)]$. *If* $\xi_3 \geq 0$ *and for all* $t \geq 0$,

$$\mathbb{P}\left(\xi_1 \geq \xi_2 + \sqrt{\xi_3 t}\right) \leq \Gamma e^{-\gamma t}, \tag{5.12}$$

then for all $t \geq 0$, *we have*

$$\mathbb{P}\left(\xi_1' \geq \xi_2' + \sqrt{\xi_3' t}\right) \leq \Gamma e^{1-\gamma t}. \tag{5.13}$$

Consider the space $\{0,1\}^n$ with uniform measure $\mathbb{P}_\varepsilon$. If $\varepsilon \in \{0,1\}^n$ and $\mathcal{A} \in \{0,1\}^n$, define the following set:

$$U_{\mathcal{A}}(\varepsilon) := \left\{(s_i)_{i \leq n} \in \{0,1\}^n, \exists \varepsilon' \in \mathcal{A}, s_i = 0 \Rightarrow \varepsilon' = \varepsilon\right\}. \tag{5.14}$$

Define the "convex hull" distance between the point ε and a set $\mathcal{A}$ as

$$f_c(\mathcal{A}, \varepsilon) := \inf\left\{|s| : s \in conv U_{\mathcal{A}}(\varepsilon)\right\},$$

where $|s|$ denotes the Euclidean norm of s. Talagrand's concentration inequalities [23] state the followings:

Proposition 5.2. *[Theorem 4.3.1 in [23]] For any* $\alpha \geq 0$, *we have*

$$\mathbb{P}_\varepsilon(f_c^2(\mathcal{A}, \varepsilon) \geq t) \leq \frac{1}{\mathbb{P}_\varepsilon(\mathcal{A})^\alpha} \exp\left(-\frac{\alpha t}{\alpha + 1}\right). \tag{5.15}$$

Proposition 5.3. *[Theorem 4.1.2 in [23]] If* $f_c^2(\mathcal{A}, \varepsilon) \leq t$, *then*

$$\forall (\lambda_i)_{i \leq n}, \exists \varepsilon' \in \mathcal{A}, \sum_{i=1}^n \lambda_i I(\varepsilon_i' \neq \varepsilon_i) \leq \sqrt{t \sum_{i=1}^n \lambda_i^2}, \tag{5.16}$$

where ε_i *and* ε_i' *are the* i*-th components of* ε *and* ε' *respectively, and* $I(x)$ *is the indicator function which returns* 1 *if* x *is true and* 0 *otherwise.*

Firstly, it is enough to prove that

$$\mathbb{P}\left(\exists f \in \mathcal{F}, S_n f \geq \Phi_n f + \sqrt{4W_n f t}\right) \leq \beta^{-\alpha} \exp\left(-\frac{\alpha t}{\alpha + 1}\right). \tag{5.17}$$

If this inequality holds, then we apply the symmetrization technique in proposition 5.1 with $\xi_1 = Q_n f$, $\xi_2 = \Phi_n f$, and $\xi_3 = 4V_n f$, and then optimize the right-hand side over α. The result yields (5.11).

Secondly, we rewrite the left-hand side of (5.17) by exploiting a fact that $\Phi_n f$, and $V_n f$ are invariant under any permutation $\pi \in \Pi$:

$$\begin{aligned} & \mathbb{P}\left(\exists f \in \mathcal{F}, S_n f \geq \Phi_n f + \sqrt{4W_n f t}\right) \\ = \; & \mathbb{P}\left(\exists f \in \mathcal{F}, R_n f \geq \Phi_n f + \sqrt{4W_n f t}\right) = \mathbb{E}\mathbb{P}_\varepsilon\left(\exists f \in \mathcal{F}, R_n f \geq \Phi_n f + \sqrt{4W_n f t}\right) \end{aligned} \tag{5.18}$$

Finally, we use Talagrand's concentration inequality on the discrete cube $\{0,1\}^n$ to complete the proof. Note that in this case we use the cube $\{-1,+1\}^n$, but no change in the inequality is needed. For a fixed pair of x and y, consider a set $\mathcal{A} = \left\{\varepsilon : \sup_{f\in\mathcal{F}}(R_n f - \Phi_n f) \leq 0\right\}$ and a set $\mathcal{A}_t = \left\{\varepsilon : f_c^2(\mathcal{A}, \varepsilon) \leq t\right\}$. Following from condition (5.10), $\mathbb{P}_\varepsilon(\mathcal{A}) \geq \beta$, we use proposition 5.2 to bound $\mathbb{P}_\varepsilon(\mathcal{A}_t)$. We obtain

$$\mathbb{P}_\varepsilon(\mathcal{A}_t) \geq 1 - \beta^{-\alpha} \exp\left(-\frac{\alpha t}{\alpha + 1}\right). \tag{5.19}$$

Next, we choose randomly $\varepsilon \in \mathcal{A}_t$ and choose $\varepsilon' \in \mathcal{A}$ according to proposition 5.3 with $\lambda_i = |a_i||f(y_i) - f(x_i)|$. Then, for any $f \in \mathcal{F}$:

$$\sum_{i=1}^n \varepsilon_i a_i (f(y_i) - f(x_i)) - \Phi_n f \leq \sum_{i=1}^n (\varepsilon_i - \varepsilon_i') a_i (f(y_i) - f(x_i)) \tag{5.20}$$

$$\leq \; 2\sum_{i=1}^n I(\varepsilon_i \neq \varepsilon_i')|a_i||f(y_i) - f(x_i)| \leq \sqrt{4t \sum_{i=1}^n a_i^2 (f(y_i) - f(x_i))^2} = \sqrt{4W_n f t}. \tag{5.21}$$

Here, (5.20) holds because $\varepsilon' \in \mathcal{A}$. This completes the proof.

In order to use Lemma 5.1, we need a functional $\Phi_n f$ that satisfies (5.10). The following lemma shows that $\Phi_n f$ does exist. In addition, (5.22) in the lemma implies that Theorem 3.1 holds.

Lemma 5.2. *Fix $x, y \in \mathcal{X}$. If (3.5) holds for some $\beta \in (0,1)$, then there exists a constant $K < \infty$ (that depends on β only) and a functional $\Phi_n f = \Phi(f, x, y)$ invariant to the permutations $\pi \in \Pi$ of (x, y) such that (5.10) holds and*

$$\mathbb{E}_y \Phi_n f \leq K \int_0^{\frac{\sqrt{V_n f}}{2m}} \sqrt{\log D(\mathcal{F}, u)} du. \tag{5.22}$$

Chứng minh. The proof is based on the standard chaining technique, which appears, for example, in Theorem 3 of [17] and in Theorem 2.5.6 and 2.14.2 of [24]. Define a set:

$$F := \{(f(x_1), \ldots, f(x_n), f(y_1), \ldots, f(y_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^{2n}, \tag{5.23}$$

and a distance function:

$$d_{x,y}(f, g) \;\; := \;\; \sqrt{\sum_{i=1}^n a_i^2 (f(y_i) - f(x_i) - g(y_i) + g(x_i))^2}. \tag{5.24}$$

Here, the weights a_i for $i = 1..n$ are incorporated into the distance function $d_{x,y}(f,g)$. One can check that $d_{x,y}(f,0) = \sqrt{W_n f}$, and that $d^2_{x,y} \leq 2(d^2_{x,0} + d^2_{0,y}) \leq 4\max\left\{d^2_{x,0}, d^2_{0,y}\right\} \leq (2qm)^2$. In addition, $D(F,u,d_{x,y}) \leq D(\mathcal{F}, u/(2m))$. Define a decreasing sequence $p_j = 2qm2^{-j}$ for all $j \geq 0$. Note that $p_0 \geq d_{x,y} \geq p_\infty = 0$.

We assume $0 \in \mathcal{F}$. Construct an increasing sequence of sets $\{0\} = F_0 \subseteq F_1 \subseteq F_2 \subseteq \ldots \subseteq F$, such that for any $g \neq h \in F_j$, $d_{x,y}(g,h) > p_j$ and for all $f \in F$ there exists $g \in F_j$ such that $d_{x,y} \leq p_j$. The cardinality of F_j is bounded by $|F_j| \leq D(F, p_j, d_{x,y}) \leq D(\mathcal{F}, q2^{-j})$. Let $r_j = D(\mathcal{F}, q2^{-j})$ for all $j \geq 0$. In addition, if $r_j = r_{j+1}$ then we construct F_j equal to F_{j+1}.

Define a sequence of projections $\pi_j : F \to F_j, j \geq 0$ in the following way. Let j be an integer such that $d_{x,y}(f,0) \in (p_{j+1}, p_j]$. For all $0 \leq k \leq j$, set $\pi_k(f) = 0$. For all $k > j$, choose $\pi_k(f)$ such that $d_{x,y}(f, \pi_k(f)) \leq p_k$. When $F_k = F_{k+1}$, let $\pi_k(f) = \pi_{k+1}(f)$. The construction leads to: $d_{x,y}(\pi_{k-1}(f), \pi_k(f)) \leq p_{k-1} + p_k = 3p_k$.

Construct another sequence $\Delta_j = \{g - h : g \in F_j, h \in F_{j-1}, d_{x,y}(g,h) \leq 3p_j\}$ for all $j \geq 1$. Let $\Delta_j = \{0\}$ if $r_j = r_{j-1}$. The cardinality of Δ_j does not exceed $|\Delta_j| \leq |F_j||F_{j-1}| \leq r_j^2$. By definition, any $f \in F$ can be represented as a sum of elements from Δ_j by the formula $f = \sum_{j\geq 1}(\pi_j(f) - \pi_{j-1}(f))$, where $\pi_j(f) - \pi_{j-1}(f) \in \Delta_j$.

For all $j \geq 1$, let $I_j = \int_{q2^{-j-1}}^{q2^{-j}} \sqrt{\log D(\mathcal{F}, u)}du$ and define the event

$$A = \bigcup_{j=1}^{\infty} \left\{ \sup_{f\in\Delta_j} \sum_{i=1}^{n} \varepsilon_i (f(y_i) - f(x_i)) \geq KI_j \right\}. \tag{5.25}$$

We are now ready to prove the lemma by bounding the occurrence of A. When A does not occur, for any $f \in F$ let j be an integer such that $\sqrt{W_n f} = d_{x,y}(f,0) \in (p_{j+1}, p_j]$,

$$\begin{aligned} R_n f &= \sum_{k\geq j+1} \sum_{i=1}^{n} \varepsilon_i a_i ((\pi_k(f) - \pi_{k-1}(f))(y_i) - (\pi_k(f) - \pi_{k-1}(f))(x_i)) & (5.26)\\ &\leq \sum_{k\geq j+1} KI_k \leq K \int_0^{q2^{-j-1}} \sqrt{\log D(\mathcal{F}, u)}du \leq K \int_0^{\frac{\sqrt{W_n f}}{2m}} \sqrt{\log D(\mathcal{F}, u)}du. & (5.27) \end{aligned}$$

If there exists $K < \infty$ such that $\mathbb{P}_\varepsilon(A) < 1 - \beta$, then by choosing

$$\Phi_n f = \frac{K}{2m} \int_0^{\sqrt{W_n f}} \sqrt{\log D(\mathcal{F}, u, d_{x,y})}du \leq K \int_0^{\frac{\sqrt{W_n f}}{2m}} \sqrt{\log D(\mathcal{F}, u)}du, \tag{5.28}$$

we get (5.10). Besides, we get (5.22) by marginalizing (5.28) over y. It remains to prove that for some $K < \infty$, $\mathbb{P}_\varepsilon(A) < 1 - \beta$. For some j such that $r_{j+1} > r_j$ and for some $f \in \Delta_j$, applying Markov's inequality, we get:

$$\mathbb{P}_\varepsilon \left(\sum_{i=1}^{n} \varepsilon_i a_i (f(y_i) - f(x_i)) \geq KI_j \right) \leq \exp\left(1 - \frac{K^2 I_j^2}{\sum_{i=1}^n a_i^2 (f(y_i) - f(x_i))^2}\right). \tag{5.29}$$

Since $d^2_{x,y}(f,0) \leq 9p_j^2$ due to $f \in \Delta_j$, and $I_j^2 \geq \log r_j q^2 2^{-2j-2}$ because $D(\mathcal{F}, u)$ is decreasing,

$$\mathbb{P}_\varepsilon \left(\sum_{i=1}^{n} \varepsilon_i a_i (f(y_i) - f(x_i)) \geq KI_j \right) \leq e r_j^{\frac{-K^2}{144m^2}}.$$

Therefore,

$$\mathbb{P}_\varepsilon(A) \le \sum_{j\ge1} |\Delta_j| e r_j^{\frac{-K^2}{144m^2}} 1_{[r_{j+1}>r_j]} \le e \sum_{j\ge1} r_j^{-\alpha} 1_{[r_{j+1}>r_j]} \le e \sum_{j\ge2} j^{-\alpha} = e(\zeta(\alpha)-1), \quad (5.30)$$

where $\alpha = \frac{K^2}{144m^2} - 2$ and $\zeta(\cdot)$ is the Riemann zeta function. To have $\mathbb{P}_\varepsilon(A) \le 1-\beta$, we can choose:

$$K = 12m\sqrt{\zeta^{-1}\left(\frac{1-\beta}{e}+1\right)+2}. \quad (5.31)$$

5.2. Proof for Theorem 3.2

This theorem is an application of Theorem 3.1 on bounding the asymmetric error. The way we derive Theorem 3.2 from Theorem 3.1 is analogous to the way Theorem 1 of [12] is derived from Corollary 3 of [16]. The differences are mainly at the terms involved. Therefore, in what follows, we will make use of some results in [10, 12] (which should not be reproduced here due to space limit) and only give a sketch of our proof. Note that the term K below may have different values during the derivation of the proof.

First of all, we obtain the following Lemma 5.3 in the same way that Theorem 6 of [12] is derived from Corollary 3 of [16], (*ie.* by upper-bounding $V_n f$ with Pf/n' since $W_n f \le P_n f/n'$). After some arrangements among the terms, we get:

Lemma 5.3. *If $\mathcal{F}' = \{\varphi_\delta(f) : f \in \mathcal{F}\}$ is a function class that satisfies* (3.5)*, then there exists an absolute constant $K < \infty$ such that for any $t > 0$ with probability at least $1 - e^{-t}$ for all $f' \in \mathcal{F}'$,*

$$Q_n f' \le K\left(\frac{1}{\sqrt{n'}}\int_0^{\sqrt{Pf'}} \sqrt{\log D(\mathcal{F}', u)}du + \sqrt{\frac{tPf'}{n'}}\right). \quad (5.32)$$

Secondly, for a fixed d, γ, we look at a layer of function $\mathcal{F}_{d,\gamma} = \{f \in \mathcal{F} : d(f;\gamma) \le d\}$. The uniform entropy of $\mathcal{F}_{d,\gamma}$ was estimated in [10],

$$\log D(\mathcal{F}_{d,\gamma}, u) \le K\left(d\log\frac{1}{u} + \left(\frac{\gamma}{u}\right)^\alpha\right). \quad (5.33)$$

Besides, define $\mathcal{F}'_{d,\gamma} := \{\varphi_\delta(yf(x)) : f \in \mathcal{F}_{d,\gamma}\}$, we have $D(\mathcal{F}'_{d,\gamma}, u) \le D(\mathcal{F}_{d,\gamma}, \delta u)$, since for any probability measure Q on $\mathcal{X}\times\mathcal{Y}$, get:

$$Q(\varphi_\delta(yf(x)) - \varphi_\delta(yg(x)))^2 \le \delta^{-2} Q(yf(x) - yg(x))^2 = \delta^{-2} Q(f(x) - g(x))^2. \quad (5.34)$$

Therefore,

$$\log D(\mathcal{F}'_{d,\gamma}, u) \le K\left(d\log\frac{1}{\delta u} + \left(\frac{\gamma}{\delta u}\right)^\alpha\right). \quad (5.35)$$

Thirdly, we apply the following well-known inequality,

$$\int_0^s \left(\log\frac{1}{u}\right)^{1/2} du \le 2s\left(\log\frac{1}{s}\right)^{1/2} \qquad \text{for } s \in [0, e^{-1}], \quad (5.36)$$

on the integral term in (5.32) after we upper-bound it with (5.35). We assume that $Pf' \geq 1/n'$, otherwise the bound of the theorem becomes trivial. We obtain for some constant $K > 0$, that

$$\int_0^{\sqrt{Pf'}} \sqrt{\log D(\mathcal{F}'_{d,\gamma}, u)} du \leq K \left(\sqrt{Pf'} \left(\frac{d}{n'} \log \frac{n'}{\delta} \right)^{1/2} + \left(\frac{\gamma}{\delta} \right)^{\alpha/2} \frac{\sqrt{Pf'}^{1-\alpha/2}}{\sqrt{n'}} \right). \quad (5.37)$$

Thus, for any $t > 0$, with probability at least $1 - e^{-t}$, for all $f \in \mathcal{F}_{d,\gamma}$, and $f'(x, y) = \varphi_\delta(yf(x))$, we have

$$\frac{Q_n f'}{\sqrt{Pf'}} \leq K \left(\sqrt{\frac{d}{n'} \log \frac{n'}{\delta}} + \left(\frac{\gamma}{\delta} \right)^{\alpha/2} \frac{\sqrt{Pf'}^{\alpha/2}}{\sqrt{n'}} + \sqrt{\frac{t}{n'}} \right). \quad (5.38)$$

Finally, we use the chaining technique proposed in the last part of the proof for Theorem 1 in [12] to complete the proof. That is, we replace t with $t' + \log \frac{Kd^2}{\delta\gamma}$ and derive a union bound over all values of γ. We get Theorem 3.2 by selecting yet another constant K large enough.

5.3. Proof for Corollary 3.1

The proof is based on the strategy that Theorem 1 in [12] was specialized in Corollary 1 in [12]. It is not reproduced here since one can follow the proof for Corollary 1 in [12]. The difference between their proof and our proof is, we replace their terms $\mathbb{P}\varphi_\delta(yf(x))$, $\mathbb{P}_n\varphi_\delta(yf(x))$, and n, with our terms Pf', $P_n f'$, and n', respectively.

REFERENCES

[1] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees.* Wadsworth International Group, 1984.

[2] Adam Cannon, James Howse, Don Hush, and Clint Scovel. Learning with the neyman–pearson and min–max criteria. Technical report, Los Alamos National Laboratory, 2002.

[3] Luc Devroye, Laszlo Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition (Stochastic Modelling and Applied Probability).* Springer, February 1997.

[4] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm, *Intl. Conf. on Machine Learning*, Bari, Italy, 1996 (148–156).

[5] H. Grabner, C. Leistner, and H. Bischof, Semi-supervised on-line boosting for robust tracking, *Europ. Conf. on Computer Vision*, (2008) pages I: 234–247.

[6] David Haussler, Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *J. Combin. Theory. Ser. A* **69** (1995) 217–232.

[7] Xinwen Hou, Cheng-Lin Liu, and Tieniu Tan, Learning boosted asymmetric classifiers for object detection, *Conf. on Computer Vision and Pattern Recognition* (New York, NY, USA) **1** (2006) 330–338.

[8] Grigoris Karakoulas and John Shawe-Taylor, Optimizing classifiers for imbalanced training sets, *Neural Information Processing Systems*, Cambridge, MA, USA, (MIT Press) 1999 (253–259).

[9] Ulrich Knoll, Gholamreza Nakhaeizadeh, and Birgit Tausend, Cost-sensitive pruning of decision trees, *Europ. Conf. on Machine Learnign* (1994) 383–386.

[10] Vladimir Koltchinskii, Dmitriy Panchenko, and Fernando Lozano, Bounding the generalization error of convex combinations of classifiers: Balancing the dimensionality and the margins, *Annals of Appl. Prob.* **13** (1) (2003) 213–252.

[11] Vladimir Koltchinskii and Dmitry Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, *The Annals of Statistics* **30** (1) (2002) 1–50.

[12] Vladimir Koltchinskii and Dmitry Panchenko, Complexities of convex combinations and bounding the generalization error in classification, *The Annals of Statistics* **33** (4) (2005) 1455–1496.

[13] Vladimir Koltchinskii, Dmitry Panchenko, and Savina Andonova, Generalization bounds for voting classifiers based on sparsity and clustering, *COLT* (2003) 492–505.

[14] Matjaz Kukar and Igor Kononenko, Cost-sensitive learning with neural networks, *13th European Conference on Artificial Intelligence* (1998) 445–449.

[15] Hamed Masnadi-Shirazi and Nuno Vasconcelos, Asymmetric boosting, *Intl. Conf. on Machine Learning* (2007) 609–619.

[16] Dmitriy Panchenko, Symmetrization approach to concentration inequalities for empirical processes, *The Annals of Probability* **31** (2003) 2068–2081.

[17] Dmitry Panchenko, Some extensions of an inequality of vapnik and chervonenkis, *Electron. Comm. Probab.* **7** (2002) 55–65.

[18] Minh-Tri Pham and Tat-Jen Cham, Online learning asymmetric boosted classifiers for object detection, *Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, 2007.

[19] Minh-Tri Pham, Viet-Dung D. Hoang, and Tat-Jen Cham, Detection with multi-exit asymmetric boosting, *Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[20] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee, Boosting the margin: A new explanation for the effectiveness of voting methods, *The Annals of Statistics* **26** (5) (1998) 1651–1686.

[21] Robert E. Schapire and Yoram Singer, Improved boosting using confidence-rated predictions, *Machine Learning* **37** (3) (1999) 297–336.

[22] C. Scott and R. Nowak, A neyman-pearson approach to statistical learning, *IEEE Transactions on Information Theory* **51** (11) 2005 (3806–3819).

[23] Michel Talagrand, Concentration of measure and isoperimetric inequalities in product spaces, *Publ. Math. l'I.H.E.S* **81** (1995) 73–205.

[24] Aad W. van der Vaart and Jon A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, November 2000.

[25] Vladimir Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[26] K. Veropoulos, N. Cristianini, and C. Campbell, Controlling the sensitivity of support vector machines, *Intl. Joint Conf. on Artificial Intelligence*, Stockholm, Sweden, 1999.

[27] Mathukumalli Vidyasagar, *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1997.

[28] Paul A. Viola and Michael J. Jones, Rapid object detection using a boosted cascade of simple features, *Conf. on Computer Vision and Pattern Recognition* **1**, Kauai, HI, USA, 2001 (511–518).

[29] Paul A. Viola and Michael J. Jones, Fast and robust classification using asymmetric adaboost and a detector cascade, *Neural Information Processing Systems*, MIT Press, Cambridge, MA, USA, 2002 (1311–1318).

[30] Bianca Zadrozny, John Langford, and Naoki Abe, Cost-sensitive learning by cost-proportionate example weighting, *IEEE International Conference on Data Mining*, 2003 (0–435).

Received on September 04, 2012
Revised on December 10, 2012