GENERALIZATION ERROR ANALYSIS FOR FDR CONTROLLED CLASSIFICATION

Clayton Scott, Gowtham Bellala

Department of Electrical
Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109

Rebecca Willett

Department of Electrical
and Computer Engineering
Duke University, Durham, NC 27708

ABSTRACT

The false discovery rate (FDR) and false nondiscovery rate (FNDR) have received considerable attention in the literature on multiple testing. These performance measures are also appropriate for classification, and in this work we develop generalization error bounds for FDR and FNDR from the perspective of statistical learning theory. Unlike more conventional classification performance measures, the empirical FDR and FNDR are not binomial random variables but rather a ratio of binomials, which introduces several challenges not addressed in conventional analyses. We develop distribution-free uniform deviation bounds and apply these, in conjunction with the Borel-Cantelli lemma, to obtain a strongly consistent learning rule.

Index Terms— Statistical learning theory, false discovery rate, supervised learning, strong consistency.

1. CONTROLLING THE FALSE DISCOVERY RATE

When learning a classifier from labeled training data, minimizing the probability of misclassification is often unsatisfactory. In a variety of applications, such as screening medical images for cancerous lesions or detecting landmines, false alarms and misses have different impacts. False detections of targets (such as lesions or landmines) are problematic because of the time, money, and other resources which are invariably wasted as a result. Missed detections, on the other hand, may result in loss of life or destruction. For this reason, a number of methods for cost-sensitive [1, 2, 3] and Neyman-Pearson [4, 5, 6] classification have been developed that allow the user to effect a tradeoff between false alarm and miss rates.

The probability of error, false alarm rate, and miss rate are all performance measures that reflect the performance of a classifier on a *single* future test point. However, it is often the case that we desire to classify *multiple* future test points. In this situation, the false alarm and miss rates may not be the most appropriate measures of performance. If a classifier has a false alarm rate of say 5%, and 1000 negative test points are observed, we expect 50 of them to be declared positive. This may be unacceptable, especially in situations where large costs are involved in investigating false alarms.

This situation is similar to the multiple testing problem in hypothesis testing. Consequently, many of the solutions to the multiple testing problem are applicable in the classification setting. The basic approach is to consider alternative measures of size and power that are better suited to multiple inference, and to design classifiers based on these new performance measures.

In this paper, we consider a measure that has been the focus of much recent work on multiple testing, the false discovery rate (FDR) [7]. Control of the FDR, i.e., the fraction of declared positives (discoveries) that are in fact negative, ensures that follow-up studies into declared positives must return a certain yield of actual positives. Such control is vital in applications where follow-up studies are time or resource consuming.

Several researchers, spurred by the seminal work of [7], have studied FDR control in the context of multiple hypothesis testing by assuming a known distribution of the observations under the null hypothesis and thresholding p-values of test statistics. It is important to note that such procedures are not applicable in the statistical learning context because we do not assume knowledge of the null distribution and must instead rely upon training data to control the FDR.

We develop basic results on the analysis of generalization error in FDR controlled classification, including uniform deviation bounds and strong consistency. The consistency result is "effectively" universal, in that it holds for all distributions for which an optimal classifier with respect to FDR/FNDR can even be reasonably defined. Unlike traditional performance probabilities, whose empirical versions are related to binomial random variables, empirical versions of FDR and FNDR are related to ratios of binomial variables. This necessitates the development of novel concentration inequalities and methods of analysis, which we illustrate for the case of countable classes of classifiers.

1.1. Problem Statement

More formally, in this paper we consider the following scenario: Let $\mathcal{X}$ be a set and $Z = (X, Y)$ be a random variable taking values in $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$. The variable X corresponds to a feature vector and Y to a class label associated with X;

$Y = 0$ corresponds to the *null hypothesis* (e.g. that no target is present) and $Y = 1$ corresponds to the *alternative hypothesis* (e.g. that a target is present). The distribution on Z is unknown and is denoted by $\mathbb{P}$. A *classifier* is a function mapping feature vectors to class labels: $h : \mathcal{X} \longrightarrow \{0, 1\}$. Let $\mathcal{H}$ denote a collection of different classifiers. Assume we make n independent and identically distributed training observations of Z, denoted $Z^n = (X_i, Y_i)_{i=1}^n$.

A "false discovery" occurs when $h(x) = 1$ but the true label is $y = 0$. Similarly, a "false nondiscovery" occurs when $h(x) = 0$ but the true label is $y = 1$. In this paper we focus on the design of classifiers for which the false discovery rate (FDR) and false nondiscovery rate (FNDR),

$$\begin{aligned} \mathcal{R}_D(h) &:= \mathbb{P}(Y = 0 \,|\, h(X) = 1) \\ \mathcal{R}_{ND}(h) &:= \mathbb{P}(Y = 1 \,|\, h(X) = 0) \end{aligned}$$

are small.

There are several slightly varying definitions for FDR/FNDR. Our definition, which is natural in the classification setting, coincides with the so-called *positive* FDR/FNDR of Storey [8, 9], so named because it can be seen to equal the expected fraction of false discoveries, conditioned on a positive number of discoveries having been made. Storey makes some decision-theoretic connections to classification, but does not consider learning from data [9].

2. UNIFORM DEVIATION BOUNDS

Define empirical analogues to the FDR and FNDR according to

$$\begin{aligned} \widehat{\mathcal{R}}_D(h) &:= \begin{cases} \frac{1}{n_D} \sum_{i=1}^n \mathbb{I}_{\{Y_i=0, h(X_i)=1\}}, & n_D > 0 \\ 0, & n_D = 0 \end{cases} \\ \widehat{\mathcal{R}}_{ND}(h) &:= \begin{cases} \frac{1}{n_{ND}} \sum_{i=1}^n \mathbb{I}_{\{Y_i=1, h(X_i)=0\}}, & n_{ND} > 0 \\ 0, & n_{ND} = 0 \end{cases} \end{aligned}$$

where $n_D = n_D(h) = \sum_{i=1}^n \mathbb{I}_{\{h(X_i)=1\}}$ and $n_{ND} = n_{ND}(h) = \sum_{i=1}^n \mathbb{I}_{\{h(X_i)=0\}}$. This section describes a uniform bound on the amount by which the empirical estimate of FDR/FNDR can deviate from the true value. Standard approaches for bounding such deviations require modification because of the random denominators in $\widehat{\mathcal{R}}_D(h)$ and $\widehat{\mathcal{R}}_{ND}(h)$.

Assume $\mathcal{H}$ is countable, and let $[\![h]\!]$ be a real valued functional on $\mathcal{H}$ such that $[\![h]\!] \geq 0$ and $\sum_{h \in \mathcal{H}} 2^{-[\![h]\!]} \leq 1$. Such a functional can be equated with a prefix code for $\mathcal{H}$, in which case $[\![h]\!]$ is the codelength associated to h.

For $0 < \delta < 1/2$, we introduce the penalty terms

$$\begin{aligned} \phi_D(h, \delta) &:= \begin{cases} \sqrt{\frac{[\![h]\!] \log 2 + \log(2/\delta)}{2 n_D(h)}}, & n_D > 0 \\ 1, & n_D = 0 \end{cases} \\ \phi_{ND}(h, \delta) &:= \begin{cases} \sqrt{\frac{[\![h]\!] \log 2 + \log(2/\delta)}{2 n_{ND}(h)}}, & n_{ND} > 0 \\ 1, & n_{ND} = 0 \end{cases} \end{aligned}$$

(All $\log$ terms are base e unless otherwise specified.) Note that these penalties depend on the training data through n_D and n_{ND}.

Theorem 1 *With probability at least $1 - \delta$ with respect to the draw of the training data,*

$$|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| \leq \phi_D(h, \delta) \tag{1}$$

for all $h \in \mathcal{H}$. Similarly, with probability at least $1 - \delta$ with respect to the draw of the training data,

$$|\mathcal{R}_{ND}(h) - \widehat{\mathcal{R}}_{ND}(h)| \leq \phi_{ND}(h, \delta) \tag{2}$$

for all $h \in \mathcal{H}$. The results are independent of the underlying probability distribution.

Because of the form of the penalty terms, the bound is looser for classifiers h that are more complex, as represented through the codelength $[\![h]\!]$, and tighter the more discoveries/nondiscoveries are made. Such bounds are key ingredients in proving consistency for penalized empirical error minimizers, as developed in the next section.

Proof: We prove the first statement, the second being analogous. For added clarity, write the penalty as $\phi_D(h, \delta, n_D)$, where

$$\phi_D(h, \delta, k) := \begin{cases} \sqrt{\frac{[\![h]\!] \log 2 + \log(2/\delta)}{2k}}, & k > 0 \\ 1, & k = 0 \end{cases}$$

Consider a fixed $h \in \mathcal{H}$. The fundamental concentration inequality underlying our bounds is Hoeffding's [10], which states that if S_k is the sum of $k > 0$ iid random variables bounded between zero and one, and $\mu = \mathbb{E}[S_k]$, then

$$\mathbb{P}(|\mu - S_k| > k\epsilon) \leq 2e^{-2k\epsilon^2}.$$

To apply Hoeffding's inequality, we need the following conditioning argument. Let $V \in \{0, 1\}^n$ be a binary indicator vector, with $V_i = \mathbb{I}_{\{h(X_i)=1\}}$. Let $\mathcal{V}_k$ denote the set of all $v \in \{0, 1\}^n$ such that $\sum_{i=1}^n v_i = k$. We may then write

$$\begin{aligned} &\mathbb{P}(|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| > \phi_D(h, \delta, n_D)) \\ &= \sum_{k=0}^n \sum_{v \in \mathcal{V}_k} \mathbb{P}(|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| > \phi_D(h, \delta, k) | V = v) \\ &\qquad \cdot \mathbb{P}(V = v) \\ &= \sum_{k=0}^n \sum_{v \in \mathcal{V}_k} \mathbb{P}(|k\mathcal{R}_D(h) - k\widehat{\mathcal{R}}_D(h)| > k\phi_D(h, \delta, k) \\ &\qquad |V = v) \cdot \mathbb{P}(V = v), \end{aligned}$$

First note that $|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| \leq \phi_D(h, \delta)$ with probability one when $n_D = 0$. We now apply Hoeffding's inequality

for each k and $v \in \mathcal{V}_k$, conditioning on $V = v$. Setting $S_k = k\widehat{\mathcal{R}}_D(h)$, we have

$$\begin{aligned}\mu &= \mathbb{E}[S_k|V=v] = k\mathbb{E}[\widehat{\mathcal{R}}_D(h)|V=v] \\ &= \mathbb{E}[\sum_{i=1}^{n} \mathbb{I}_{\{Y_i=0,h(X_i)=1\}}|V=v] \\ &= \mathbb{E}[\sum_{i:v_i=1} \mathbb{I}_{\{Y_i=0\}}|V=v] \\ &= k\mathbb{P}(Y=0|h(X)=1) = k\mathcal{R}_D(h),\end{aligned}$$

where in the last step we use independence of the realizations. Applying Hoeffding's inequality conditioned on $V = v \in \mathcal{V}_k$ yields

$$\begin{aligned}&\mathbb{P}(|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| > \phi_D(h,\delta,n_D)) \\ &\leq \sum_{k=1}^{n}\sum_{v\in\mathcal{V}_k} 2e^{-2k\phi_D^2(h,\delta,k)}\mathbb{P}(V=v) \\ &\leq \sum_{k=1}^{n}\sum_{v\in\mathcal{V}_k} \delta 2^{-[\![h]\!]}\mathbb{P}(V=v) \\ &= \delta 2^{-[\![h]\!]}(1-\mathbb{P}(\textstyle\sum V_i = 0)) \leq \delta 2^{-[\![h]\!]}.\end{aligned}$$

The result now follows by applying the union bound over all $h \in \mathcal{H}$. ■

The technique of conditioning on the random denominator of a ratio of binomials has also been applied in others settings [11, 5].

3. MEASURING PERFORMANCE

We would like to be able to make FDR/FNDR related guarantees about how a data-based classifier $\widehat{h}$ performs. For this, we need to specify a performance measure or optimality criterion that incorporates both FDR and FNDR quantities simultaneously. One possibility is to specify a number $0 < \alpha < 1$ and seek the classifier such that $\mathcal{R}_{ND}(h)$ is minimal while $\mathcal{R}_D(h) \leq \alpha$ with high probability. Another is to specify a constant $\lambda > 0$ reflecting the relative cost of FDR to FNDR, and minimize

$$\mathcal{E}_\lambda(h) := \mathcal{R}_{ND}(h) + \lambda\mathcal{R}_D(h).$$

The uniform deviation bounds developed in the previous section allow us to analyze both criteria, but for now we focus on the latter. In particular, the uniform deviation bounds immediately imply

Corollary 1 *With probability at least $1 - 2\delta$ with respect to the draw of the training data,*

$$\mathcal{E}_\lambda(h) \leq \widehat{\mathcal{R}}_{ND}(h) + \phi_{ND}(h,\delta_n) + \lambda[\widehat{\mathcal{R}}_D(h) + \phi_D(h,\delta_n)]$$

for all $h \in \mathcal{H}$.

4. STRONG CONSISTENCY

Denote the globally optimal value of the performance measure by

$$\mathcal{E}_\lambda^* := \inf_h \mathcal{E}_\lambda(h).$$

We seek a learning rule $\widehat{h}_{\lambda,n}$ such that $\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) \to \mathcal{E}_\lambda^*$ almost surely. Thus let $\{\mathcal{H}_k\}_{k\geq 1}$ be a family of finite classes of classifiers with universal approximation capability. That is, assume

$$\lim_{k\to\infty} \inf_{h\in\mathcal{H}_k} \mathcal{E}_\lambda(h) = \mathcal{E}_\lambda^*.$$

For example, if $\mathcal{X} = [0,1]^d$, we may take $\mathcal{H}_k$ to be the collection of histogram classifiers based on a binwidth of $1/k$. Furthermore, take $[\![h]\!] = \log_2|\mathcal{H}_k|$ for $h \in \mathcal{H}_k$, where $|\mathcal{H}_k|$ is the cardinality of $\mathcal{H}_k$. For histograms, we have $|\mathcal{H}_k| = 2^{k^d}$ and hence $[\![h]\!] = k^d$.

The bound of Corollary 1 suggests bound minimization as a strategy for selecting a classifier empirically. However, rather than minimizing over all possible classifiers in some $\mathcal{H}_k$, we first discard those classifiers whose empirical numbers of discoveries or nondiscoveries are too small. In these cases, the penalties are possibly quite large, and we are unable to obtain tight concentrations of empirical FDR/FNDR measures around their population versions. However, as we will see, the criterion for exclusion is less strict as n increases, and ultimately we are able to obtain strong consistency for essentially all meaningful distributions. This aspect of the analysis is one element that seems unique to FDR/FNDR criteria compared to traditional performance measures.

Formally, set $\delta_n = 1/n^2$ and define

$$\widehat{\mathcal{H}}_n := \{h \in \mathcal{H}_{k_n} : \frac{n_{ND}(h)}{n} \geq p_n - \epsilon_n, \frac{n_D(h)}{n} \geq p_n - \epsilon_n\},$$

where $p_n := (\log n)^{-1}$ and

$$\epsilon_n := \sqrt{\frac{\log|\mathcal{H}_{k_n}| + \log(2/\delta_n)}{2n}}.$$

Here k_n is such that $k_n \to \infty$ as $n \to \infty$ and $\log|\mathcal{H}_{k_n}| = o(n/\log^2 n)$. This condition on k_n ensures that $\epsilon_n = o(p_n)$. For histograms, $\log|\mathcal{H}_{k_n}| = k_n^d \log 2$, and thus the assumed conditions on the growth of k_n are essentially the same as for consistency of histograms in other problems [12]. The quantities p_n and ϵ_n are justified in the proof of Theorem 2 below; however, note that both quantities vanish as n grows. Also note that other (more rapidly decaying) choices for p_n will suffice, provided $\epsilon_n = o(p_n)$ still holds.

Denote the bound of Corollary 1 by

$$\widehat{\mathcal{E}}_\lambda(h) := \widehat{\mathcal{R}}_{ND}(h) + \phi_{ND}(h,\delta_n) + \lambda[\widehat{\mathcal{R}}_D(h) + \phi_D(h,\delta_n)],$$

and define the classification rule

$$\widehat{h}_{\lambda,n} := \arg\min_{h\in\widehat{\mathcal{H}}_n} \widehat{\mathcal{E}}_\lambda(h).$$

We impose one condition on the underlying distribution:

Assumption A: If $\{h_k\}$ is a sequence of classifiers such that $\mathcal{E}_\lambda(h_k) \to \mathcal{E}^*_\lambda$, as $k \to \infty$, then there exists $p > 0$ such that, for all k sufficiently large, $P(h_k(X) = 0) > p$ and $P(h_k(X) = 1) > p$.

This assumption is extremely weak. Furthermore, the FDR/FNDR criteria are hardly justified when it is violated, for then we may find a subsequence $\{h_{k_j}\}$ such that either $\mathbb{P}(h_{k_j}(X) = 0) \to 0$ or $\mathbb{P}(h_{k_j}(X) = 1) \to 0$, meaning some limiting classifier h^*_λ has $\mathbb{P}(h^*_\lambda(X) = 0)$ or $\mathbb{P}(h^*_\lambda(X) = 1)$, in which case either FDR or FNDR is undefined, and one label occurs with probability one. Succinctly, Assumption A is valid for all distributions for which FDR/FNDR makes sense.

Theorem 2 *Under Assumption A,*

$$\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) \to \mathcal{E}^*_\lambda$$

almost surely. That is, $\widehat{h}_{\lambda,n}$ is strongly consistent.

Proof: By the Borel-Cantelli lemma [13, 12], it suffices to show that for each $\epsilon > 0$

$$\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \mathcal{E}^*_\lambda \geq \epsilon) < \infty.$$

Let $\epsilon > 0$. Define the events

$$\begin{aligned}
\Omega^n &:= \{Z^n : \mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \mathcal{E}^*_\lambda \geq \epsilon\} \\
\Omega^n_1 &:= \{Z^n : \mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \inf_{h \in \widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h) \geq \frac{\epsilon}{2}\} \\
\Omega^n_2 &:= \{Z^n : \inf_{h \in \widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h) - \mathcal{E}^*_\lambda \geq \frac{\epsilon}{2}\}
\end{aligned}$$

Since $\Omega^n \subset \Omega^n_1 \bigcup \Omega^n_2$, we have

$$\sum_{n=1}^{\infty} \mathbb{P}(\Omega^n) \leq \sum_{n=1}^{\infty} \mathbb{P}(\Omega^n_1) + \sum_{n=1}^{\infty} \mathbb{P}(\Omega^n_2).$$

We consider the two terms individually and show that each of them is finite. To bound the first term we use the following lemma.

Lemma 1

$$\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \inf_{h \in \widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h) \leq 2 \sup_{h \in \widehat{\mathcal{H}}_n} |\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h)|$$

Proof: Write

$$\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \widehat{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) + \widehat{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) - \inf_{h \in \widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h).$$

The first term can be bounded as follows:

$$\begin{aligned}
\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \widehat{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) &\leq |\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \widehat{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n})| \\
&\leq \sup_{h \in \widehat{\mathcal{H}}_n} |\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h)|.
\end{aligned}$$

Now, by the definition of $\widehat{h}_{\lambda,n}$, for any $h \in \widehat{\mathcal{H}}_n$,

$$\begin{aligned}
\widehat{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) - \mathcal{E}_\lambda(h) &\leq \widehat{\mathcal{E}}_\lambda(h) - \mathcal{E}_\lambda(h) \\
&\leq |\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h)| \\
&\leq \sup_{h \in \widehat{\mathcal{H}}_n} |\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h)|
\end{aligned}$$

which implies

$$\widehat{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) - \sup_{h \in \widehat{\mathcal{H}}_n} |\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h)| \leq \inf_{h \in \widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h)$$

and the result follows. ■

Define the events

$$\begin{aligned}
\Omega^n_{11} &:= \{Z^n : \sup_{h \in \widehat{\mathcal{H}}_n} |\mathcal{R}_{ND}(h) - \widehat{\mathcal{R}}_{ND}(h)| \geq \frac{\epsilon}{16}\} \\
\Omega^n_{12} &:= \{Z^n : \sup_{h \in \widehat{\mathcal{H}}_n} |\mathcal{R}_{D}(h) - \widehat{\mathcal{R}}_{D}(h)| \geq \frac{\epsilon}{16\lambda}\} \\
\Omega^n_{13} &:= \{Z^n : \sup_{h \in \widehat{\mathcal{H}}_n} |\phi_{ND}(h, \delta_n)| \geq \frac{\epsilon}{16}\} \\
\Omega^n_{14} &:= \{Z^n : \sup_{h \in \widehat{\mathcal{H}}_n} |\phi_{D}(h, \delta_n)| \geq \frac{\epsilon}{16\lambda}\}
\end{aligned}$$

From Lemma 1 it follows that

$$\Omega^n_1 \subset \bigcup_{i=1}^{4} \Omega^n_{1i}$$

and hence it suffices to show

$$\sum_{n=1}^{\infty} \mathbb{P}(\Omega^n_{1i})$$

is finite for each $i = 1, 2, 3, 4$. We shall consider Ω_{11} and Ω_{13}, the other two cases following similarly.

For $h \in \widehat{\mathcal{H}}_n$ we have $n_{ND}(h)/n \geq p_n - \epsilon_n \geq \frac{1}{2} p_n$ for n sufficiently large, and therefore

$$\begin{aligned}
\phi_{ND}(h, \delta_n) &= \sqrt{\frac{\log |\mathcal{H}_{k_n}| + \log(2n^2)}{2 n_{ND}(h)}} \\
&\leq \sqrt{(\log |\mathcal{H}_{k_n}| + \log(2n^2)) \frac{\log n}{n}} < \frac{\epsilon}{16}
\end{aligned}$$

for $n \geq N_1$, for some N_1 sufficiently large. Then

$$\sum_{n=1}^{\infty} \mathbb{P}(\Omega^n_{13}) \leq N_1.$$

Furthermore, by the uniform deviation bound,

$$\sum_{n=1}^{\infty} \mathbb{P}(\Omega^n_{11}) \leq N_1 + \sum_{n=N_1+1}^{\infty} \frac{1}{n^2} < \infty.$$

Let us now proceed to the event Ω_2^n. Introduce

$$\mathcal{H}_n^{'} := \{h \in \mathcal{H}_{k_n} : \mathbb{P}(h(X)=0) \geq p_n, \mathbb{P}(h(X)=1) \geq p_n\}$$

and the events

$$\begin{aligned}
\Omega_{21}^n &:= \{Z^n : \inf_{h \in \widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h) - \inf_{h \in \mathcal{H}_n^{'}} \mathcal{E}_\lambda(h) \geq \frac{\epsilon}{6}\} \\
\Omega_{22}^n &:= \{Z^n : \inf_{h \in \mathcal{H}_n^{'}} \mathcal{E}_\lambda(h) - \mathcal{E}_\lambda(h_{\lambda,k_n}^*) \geq \frac{\epsilon}{6}\} \\
\Omega_{23}^n &:= \{Z^n : \mathcal{E}_\lambda(h_{\lambda,k_n}^*) - \mathcal{E}_\lambda^* \geq \frac{\epsilon}{6}\}
\end{aligned}$$

Here h_{λ,k_n}^* is the optimal classifier from $\mathcal{H}_{k_n}$. Clearly, $\Omega_2^n \subset \bigcup_{i=1}^3 \Omega_{2i}^n$, hence by the union bound it suffices to show

$$\sum_{n=1}^{\infty} \mathbb{P}(\Omega_{2i}^n) < \infty$$

for each $i = 1, 2, 3$. The third case, Ω_{23}, follows immediately by the universal approximation capability of $\{\mathcal{H}_k\}_{k \geq 1}$. The second case, Ω_{22}, follows by Assumption A, since p_n becomes arbitrarily small. Now take Ω_{21}. For each h, we have that $n_{ND}(h)$ and $n_D(h)$ are binomial random variables with parameters $\mathbb{P}(h(X)=0)$ and $\mathbb{P}(h(X)=1)$, respectively, and n. Therefore, by Chernoff's bound and the union bound, we have that

$$\left| \frac{n_{ND}(h)}{n} - \mathbb{P}(h(X)=0) \right| \leq \epsilon_n \ \ \forall h \in \mathcal{H}_{k_n}$$

and

$$\left| \frac{n_D(h)}{n} - \mathbb{P}(h(X)=1) \right| \leq \epsilon_n \ \ \forall h \in \mathcal{H}_{k_n}$$

hold together with probability at least $1-2\delta_n$. Therefore, with probability at least $1-2\delta_n$, we have that $\mathcal{H}_n^{'} \subseteq \widehat{\mathcal{H}}_n$, which implies $\inf_{h \in \widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h) - \inf_{h \in \mathcal{H}_n^{'}} \mathcal{E}_\lambda(h) \leq 0$. Therefore,

$$\sum_{n=1}^{\infty} \mathbb{P}(\Omega_{21}^n) \leq \sum_{n=1}^{\infty} \frac{2}{n^2} < \infty.$$

■

5. DISCUSSION

This paper demonstrates that FDR and FNDR control is possible in the context of statistical learning theory, where the distribution of (X, Y) is unknown except through training data. We develop empirical estimates of these quantities and derive uniform deviation bounds which assess the closeness of these empirical estimates to the true FDR and FNDR. Unlike most other performance measures in statistical learning theory, which are related to binomial random variables, the FDR and FNDR measures are related to ratios of binomial random variables and necessitate the development of novel bounding techniques. These bounds are then used to define a classification rule which is strongly consistent subject to an extremely mild conditiony on the distribution (X, Y).

6. REFERENCES

[1] F. R. Bach, D. Heckerman, and E. Horvitz, "Considering cost asymmetry in learning classifiers," *J. Machine Learning Research*, pp. 1713–1741, 2006.

[2] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, Washington, USA, 2001, pp. 973–978.

[3] B. Zadrozny, J. Langford, and N. Abe, "Cost sensitive learning by cost-proportionate example weighting," in *Proceedings of the 3rd International Conference on Data Mining*, Melbourne, FA, USA, 2003, IEEE Computer Society Press.

[4] A. Cannon, J. Howse, D. Hush, and C. Scovel, "Learning with the Neyman-Pearson and min-max criteria," Tech. Rep. LA-UR 02-2951, Los Alamos National Laboratory, 2002.

[5] C. Scott and R. Nowak, "A Neyman-Pearson approach to statistical learning," *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 3806–3819, 2005.

[6] C. Scott, "Performance measures for Neyman-Pearson classification," *IEEE Trans. Inform. Theory*, vol. 53, no. 8, 2007, to appear.

[7] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Statist. Soc B*, vol. 57, no. 1, pp. 289–300, 1995.

[8] J.D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B*, vol. 64, pp. 479–498, 2002.

[9] J.D. Storey, "The positive false discovery rate: A Bayesian interpretation of the q-value," *Annals of Statistics*, vol. 31:6, pp. 2013–2035, 2003.

[10] C. McDiarmid, "Concentration," in *Probabilistic Methods for Algorithmic Discrete Mathematics*, Berlin, 1998, pp. 195–248, Springer.

[11] Y. Mansour and D. McAllester, "Generalization bounds for decision trees," in *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, N. Cesa-Bianchi and S. Goldman, Eds., Palo Alto, CA, 2000, pp. 69–74.

[12] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[13] R. Durrett, *Probability: Theory and Examples*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1991.