Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Semi-supervised novelty detection

G. Blanchard[1],G. Lee[2],C. Scott[2]

submitted: December 11, 2009

[1] Weierstrass Institute for
Applied Analysis and Stochastics
Mohrenstrasse 39
10117 Berlin, Germany
E-Mail: blanchard@wias-berlin.de

[2] Department of Electrical Engineering
and Computer Science
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122, USA
E-Mail: {gyemin,cscott}@eecs.umich.edu

No. 1471
Berlin 2009

2010 *Mathematics Subject Classification.* 62H30,62H15.

Key words and phrases. Semi-supervised learning, novelty detection, Neyman-Pearson classification, learning reduction, two-sample problem, multiple testing.

C. Scott was supported in part by NSF Award No. 0830490. G. Blanchard acknowledges partial support of the PASCAL2 network of excellence of the European community, FP7 ICT-216886.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

A common setting for novelty detection assumes that labeled examples from the nominal class are available, but that labeled examples of novelties are unavailable. The standard (inductive) approach is to declare novelties where the nominal density is low, which reduces the problem to density level set estimation. In this paper, we consider the setting where an unlabeled and possibly contaminated sample is also available at learning time. We argue that novelty detection in this semi-supervised setting is naturally solved by a general reduction to a binary classification problem. In particular, a detector with a desired false positive rate can be achieved through a reduction to Neyman-Pearson classification. Unlike the inductive approach, semi-supervised novelty detection (SSND) yields detectors that are optimal (e.g., statistically consistent) regardless of the distribution on novelties. Therefore, in novelty detection, unlabeled data have a substantial impact on the theoretical properties of the decision rule. We validate the practical utility of SSND with an extensive experimental study.

We also show that SSND provides distribution-free, learning-theoretic solutions to two well known problems in hypothesis testing. First, our results provide a general solution to the general two-sample problem, that is, the problem of determining whether two random samples arise from the same distribution. Second, a specialization of SSND coincides with the standard *p*-value approach to multiple testing under the so-called random effects model. Unlike standard rejection regions based on thresholded *p*-values, the general SSND framework allows for adaptation to arbitrary alternative distributions.

1 Introduction

Several recent works in the machine learning literature have addressed the issue of novelty detection. The basic task is to build a decision rule that distinguishes *nominal* from *novel* patterns. The learner is given a random sample $x_1, \ldots, x_m \in \mathcal{X}$ of nominal patterns, obtained, for example, from a controlled experiment or an expert. Labeled examples of novelties, however, are not available. The standard approach has been to estimate a level set of the nominal density [Schölkopf et al., 2001, Steinwart et al., 2005, Scott and Nowak, 2006, Vert and Vert, 2006, El-Yaniv and Nisenson, 2007, Hero, 2007], and to declare test points outside the estimated level set to be novelties. We refer to this approach as *inductive* novelty detection.

In this paper we incorporate unlabeled data into novelty detection, and argue that this framework offers substantial advantages over the inductive approach. In particular, we assume that in addition to the nominal data, we also have access to an *unlabeled* sample

$x_{m+1}, \ldots, x_{m+n}$ consisting potentially of both nominal and novel data. We assume that each x_i, $i = m+1, \ldots, m+n$ is paired with an unobserved label $y_i \in \{0,1\}$ indicating its status as nominal ($y_i = 0$) or novel ($y_i = 1$), and that $(x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)$ are realizations of the random pair (X,Y) with joint distribution P_{XY}. The marginal distribution of an unlabeled pattern X is the contamination model

$$X \sim P_X = (1-\pi)P_0 + \pi P_1,$$

where P_y, $y = 0,1$, is the conditional distribution of $X|Y = y$, and $\pi = P_{XY}(Y = 1)$ is the a priori probability of a novelty. Similarly, we assume $x_1, \ldots, x_m$ are realizations of P_0. We assume no knowledge of P_X, P_0, P_1, or π, although in Section 6 (where we want to estimate the proportion π) we do impose a natural condition on P_1 which ensures identifiability of π.

We take as our objective to build a decision rule with a small false negative rate subject to a fixed constraint α on the false positive rate. Our emphasis here is on *semi-supervised* novelty detection (SSND), where the goal is to construct a general detector that could classify an arbitrary test point. This general detector can of course be applied in the *transductive* setting, where the goal is to predict the labels $y_{m+1}, \ldots, y_{m+n}$ associated with the unlabeled data. Our results extend in a natural way to this setting.

Our basic contribution is to develop a general solution to SSND by reducing it to Neyman-Pearson (NP) classification, which is the problem of binary classification subject to a user-specified constraint on the false positive rate. In particular, we argue that SSND can be addressed by applying a NP classification algorithm, treating the nominal and unlabeled samples as the two classes. We argue that our approach can effectively adapt to any novelty distribution P_1, in contrast to the inductive approach which is only optimal in certain extremely unlikely scenarios. Our learning reduction allows us to import existing statistical performance guarantees for Neyman-Pearson classification [Cannon et al., 2002, Scott and Nowak, 2005] and thereby deduce generalization error bounds, consistency, and rates of convergence for novelty detection. In addition to these theoretical properties, the reduction to NP classification has practical advantages, in that it allows essentially any algorithm for NP classification to be applied to SSND.

SSND is particularly suited to situations where the novelties occupy regions where the nominal density is high. If a single novelty lies in a region of high nominal density, it will appear nominal. However, if many such novelties are present, the unlabeled data will be more concentrated than one would expect from just the nominal component, and the presence of novelties can be detected. SSND may also be thought of as semi-supervised classification in the setting where labels from one class are difficult to obtain (see discussion of LPUE below). We emphasize that we do not assume that novelties are rare, i.e., that π is very small, as in anomaly detection. However, SSND is applicable to anomaly detection provided n is sufficiently large.

We also discuss estimation of π and the special case of $\pi = 0$, which is not treated in our initial analysis. We present a hybrid approach that automatically reverts to the inductive approach when $\pi = 0$, while preserving the benefits of the NP reduction when $\pi > 0$. In addition, we describe a distribution-free one-sided confidence interval for π, consistent estimation of π, and testing for $\pi = 0$, which amounts to a general version of the two-sample problem in statistics. We also discuss connections to multiple testing, where we show that SSND generalizes a standard approach to multiple testing, based on thresholding p-values, under the common "random effects" model. Whereas the p-value approach is optimal only under strong assumptions on the alternative distribution, SSND can optimally adapt to arbitrary alternatives.

The paper is structured as follows. After reviewing related work in the next section, we present the general learning reduction to NP classification in Section 3, and apply this reduction in Section 4 to deduce statistical performance guarantees for SSND. Section 5 presents our hybrid approach, while Section 6 applies learning-theoretic principles to the estimation of π. Connections to multiple testing are developed in Section 7. Experiments are presented in Section 8, while conclusions are discussed in the final section. Shorter proofs are presented in the main text, and longer proofs appear in the first appendix.

2 Related work

Inductive novelty detection: Described in the introduction, this problem is also known as one-class classification [Schölkopf et al., 2001] or learning for only positive (or only negative) examples. The standard approach has been to assume that novelties are outliers with respect to the nominal distribution, and to build a novelty detector by estimating a level set of the nominal density [Scott and Nowak, 2006, Vert and Vert, 2006, El-Yaniv and Nisenson, 2007, Hero, 2007]. As we discuss below, density level set estimation is equivalent to assuming that novelties are uniformly distributed. Therefore these methods can perform arbitrarily poorly (when P_1 is far from uniform, and still has significant overlap with P_0). In Steinwart et al. [2005], inductive novelty detection is reduced to classification of P_0 against P_1 , wherein P_1 can be arbitrary. However an i.i.d. sample from P_1 is assumed to be available in addition to the nominal data. In contrast, the semi-supervised approach optimally adapts to P_1, where only an unlabeled contaminated sample is available besides the nominal data. In addition, we address estimation and testing of the proportion of novelties.

Classification with unlabeled data: In transductive and semi-supervised classification, labeled training data $\{(x_i,y_i)\}_{i=1}^m$ from *both* classes are given. The setting proposed here is a special case where training data from only one class are available. In two-class problems, unlabeled data typically have at best a slight affect on constants, finite sample bounds, and rates [Rigollet, 2007, Lafferty and Wasserman, 2008, Ben-David et al., 2008, Singh et al., 2009], and are not needed for consistency. In contrast, we argue that for novelty detection, unlabeled data are essential for these desirable theoretical properties to hold.

Learning from positive and unlabeled examples: Classification of an unlabeled sample given data from one class has been addressed previously, but with certain key differences from our work. This body of work is often termed learning from "positive" and unlabeled examples (LPUE), although in our context we tend to think of nominal examples as negative. Terminology aside, a number of algorithms have been developed which proceed roughly as follows: First, identify a reliable set of negative examples in the unlabeled data. Second, iteratively apply a classification algorithm to the unlabeled data until a stable labeling is reached. Several such algorithms are reviewed in Zhang and Lee [2005], but they tend to be heuristic in nature and sensitive to the initial choice of negative examples.

A theoretical analysis of LPUE is provided by Denis [1998], Denis et al. [2005] from the point of view of computer-theoretic PAC learnable classes in polynomial time. While some ideas are common with the present work (such as classifying the nominal sample against the contaminated sample as a proxy for the ultimate goal), our point of view is relatively different and based on statistical learning theory. In particular, our input space can be non-discrete and we assume the distributions P_0 and P_1 can overlap, which leads us to use the NP classification setting and study universal consistency properties.

We highlight here one strand of LPUE research having particular relevance to our own. The idea of reducing LPUE to a binary classification problem, by viewing the positive data as one class and the unlabeled data as the other, has been treated by Zhang and Lee [2005], Liu et al. [2002], Lee and Liu [2003], Liu et al. [2003]. Most notably, Liu et al. [2002] provide sample complexity bounds for VC classes for the learning rule that minimizes the number of false negatives while controlling the proportion of false positives at a certain level. Our approach extends theirs in several respects. First, Liu et al. [2002] does not consider approximation error or consistency, nor do the bounds established there imply consistency. In contrast, we present a general reduction that is not specific to any particular learning algorithm, and can be used to deduce consistency or rates of convergence. Our work also makes several contributions not addressed previously in the LPUE literature, including our results relating to the case $\pi = 0$, to the estimation of π, and to multiple testing.

We also note recent work by A. Smola [2009] described as *relative novelty detection*. This work is presented as an extension of standard one-class classification to a setting where a reference measure (indicating regions where novelties are more likely) is known through a sample. In practice, the authors take this sample to be a contaminated sample consisting of both nominal and novel measurements, so the setting is the same as ours. The emphasis in this work is primarily on a new kernel method, whereas our work features a general learning reduction and learning theoretic analysis.

Multiple testing: The multiple testing problem is also concerned with the simultaneous detection of many potentially abnormal measurements (viewed as rejected null hypotheses). In Section 7, we discuss in detail the relation of our contamination model to the *random effects model*, a standard model in multiple testing. We show how SSND is, in several respects, a generalization of that model, and in particular includes directly several different extensions proposed in the recent multiple testing literature. The SSND model, and the results presented in this paper, are thus of particular relevance for multiple testing as well, and suggest an interesting point of view to this domain. In particular, through a reduction to classification, we introduce broad connections to statistical learning theory.

3 The fundamental reduction

To begin, we first consider the population version of the problem, where the distributions are known completely. Recall that $P_X = (1-\pi)P_0 + \pi P_1$ is the distribution of unlabeled test points. Adopting a hypothesis testing perspective, we argue that the optimal tests for $H_0 : X \sim P_0$ vs. $H_1 : X \sim P_1$ are identical to the optimal tests for $H_0 : X \sim P_0$ vs. $H_X : X \sim P_X$. The former are the tests we would like to have, and the latter are tests we can estimate by treating the nominal and unlabeled samples as labeled training data for a binary classification problem.

To offer some intuition, we first assume that P_y has density h_y, $y = 0,1$. According to the Neyman-Pearson lemma [Lehmann, 1986], the optimal test with size (false positive rate) α for $H_0 : X \sim P_0$ vs. $H_1 : X \sim P_1$ is given by thresholding the likelihood ratio $h_1(x)/h_0(x)$ at an appropriate value. Similarly, letting $h_X = (1-\pi)h_0 + \pi h_1$ denote the density of P_X, the optimal tests for $H_0 : X \sim P_0$ vs. $H_X : X \sim P_X$ are given by thresholding $h_X(x)/h_0(x)$. Now notice

$$\frac{h_X(x)}{h_0(x)} = (1-\pi) + \pi \frac{h_1(x)}{h_0(x)}.$$

Thus, the likelihood ratios are related by a simple monotone transformation, provided $\pi > 0$. Furthermore, the two problems have the same null hypothesis. Therefore, by the theory of uniformly most powerful tests [Lehmann, 1986], the optimal test of size α for one problem is also optimal, *with the same size* α, for the other problem. In other words, we can discriminate P_0 from P_1 by discriminating between the nominal and unlabeled distributions. Note the above argument does not require knowledge of π other than $\pi > 0$.

The hypothesis testing perspective also sheds light on the inductive approach. In particular, estimating the nominal level set $\{x : h_0(x) \geq \lambda\}$ is equivalent to thresholding $1/h_0(x)$ at $1/\lambda$. Thus, the density level set is an optimal decision rule provided h_1 is constant on the support of h_0. This assumption that P_1 is uniform on the support of P_0 is therefore implicitly adopted by a majority of works on novelty detection.

We now drop the requirement that P_0 and P_1 have densities. Let $f : \mathbb{R}^d \to \{0,1\}$ denote a classifier. For $y = 0,1$, let

$$R_y(f) := P_y(f(X) \neq y)$$

denote the false positive rate (FPR) and false negative rate (FNR) of f, respectively. For greater generality, suppose we restrict our attention to some fixed set of classifiers $\mathcal{F}$ (possibly the set of all classifiers). The optimal FNR for a classifier of the class $\mathcal{F}$ with FPR $\leq \alpha$, $0 \leq \alpha \leq 1$, is

$$\begin{aligned} R^*_{1,\alpha}(\mathcal{F}) \quad := \quad & \inf_{f \in \mathcal{F}} \; R_1(f) \\ & \text{s.t.} \;\; R_0(f) \leq \alpha. \end{aligned} \tag{1}$$

Similarly, introduce

$$\begin{aligned} R_X(f) \quad &:= \quad P_X(f(X) = 0) \\ &= \quad \pi R_1(f) + (1-\pi)(1 - R_0(f)) \end{aligned}$$

and let

$$\begin{aligned} R^*_{X,\alpha}(\mathcal{F}) \;:=\; & \inf_{f\in\mathcal{F}} \; R_X(f) \\ & \text{s.t. } R_0(f) \le \alpha. \end{aligned} \tag{2}$$

In this paper we will always assume the following property (involving $\mathcal{F}$, P_0 and P_1) holds:

(A) For any $\alpha \in (0,1)$, there exists $f^* \in \mathcal{F}$ such that $R_0(f^*) = \alpha$ and $R_1(f^*) = R^*_{1,\alpha}(\mathcal{F})$.

It is possible to ensure that this assumption is satisfied if one extends the class $\mathcal{F}$ to a larger class containing randomized classifiers obtained by convex combination of classifiers of the original class. This construction is standard in the receiver operating characteristic (ROC) literature. Some basic results on this topic are recalled in Appendix B in relation to the above assumption.

By the following result, the optimal classifiers for problems (1) and (2) are the same. Furthermore, one direction of this equivalence also holds in an approximate sense. In particular, approximate solutions to $X \sim P_0$ vs. $X \sim P_X$ translate to approximate solutions for $X \sim P_0$ vs. $X \sim P_1$. The following theorem constitutes our main *learning reduction* in the sense of Beygelzimer et al. [2005]:

Theorem 1 *Assume property* **(A)** *is satisfied. Consider any* α, $0 \le \alpha \le 1$, *and assume* $\pi > 0$. *If* $R_1(f) = R^*_{1,\alpha}(\mathcal{F})$ *and* $R_0(f) = \alpha$, *then* $R_X(f) = R^*_{X,\alpha}(\mathcal{F})$. *Conversely, if* $R_X(f) = R^*_{X,\alpha}(\mathcal{F})$, *and* $R_0(f) \le \alpha$ *then* $R_1(f) = R^*_{1,\alpha}(\mathcal{F})$ *and* $R_0(f) = \alpha$.

More generally, let $L_{1,\alpha}(f,\mathcal{F}) = R_1(f) - R^*_{1,\alpha}(\mathcal{F})$ *and* $L_{X,\alpha}(f,\mathcal{F}) = R_X(f) - R^*_{X,\alpha}(\mathcal{F})$ *denote the excess losses (regrets) for the two problems, and assume* $\pi > 0$. *If* $R_0(f) \le \alpha + \varepsilon$, *then*

$$L_{1,\alpha}(f,\mathcal{F}) \le \pi^{-1}(L_{X,\alpha}(f,\mathcal{F}) + (1-\pi)\varepsilon).$$

Proof . For any classifier f, we have the relation $R_X(f) = (1-\pi)(1-R_0(f)) + \pi R_1(f)$. First consider $f \in \mathcal{F}$ such that $R_1(f) = R^*_{1,\alpha}(\mathcal{F})$ and $R_0(f) = \alpha$, but assume $R_X(f) > R^*_X(\mathcal{F})$. Let $f' \in \mathcal{F}$ such that $R_X(f') < R_X(f)$ and $R_0(f') \le \alpha$. Then since $\pi > 0$,

$$\begin{aligned} R_1(f') &= \pi^{-1}\left(R_X(f') - (1-\pi)(1-R_0(f'))\right) \\ &< \pi^{-1}\left(R_X(f) - (1-\pi)(1-\alpha)\right) \\ &= R_1(f), \end{aligned}$$

contradicting minimality of $R_1(f)$.

Conversely, let $f \in \mathcal{F}$ be such that $R_X(f) = R^*_{X,\alpha}(\mathcal{F})$ and $R_0(f) \le \alpha$, but assume $R_1(f) > R^*_{1,\alpha}(\mathcal{F})$ or $R_0(f) < \alpha$. Let f' be such that $R_0(f') = \alpha$ and $R_1(f') = R^*_1(\mathcal{F})$ (whose existence is ensured by assumption **(A)**). Then

$$\begin{aligned} R_X(f') &= (1-\pi)(1-\alpha) + \pi R_1(f') \\ &< (1-\pi)(1-R_0(f)) + \pi R_1(f) \\ &= R_X(f) \\ &= R^*_{X,\alpha}(\mathcal{F}) \end{aligned}$$

contradicting minimality of $R_X(f)$. To prove the final statement, first note that we established $R^*_{X,\alpha}(\mathcal{F}) = \pi R^*_{1,\alpha}(\mathcal{F}) + (1-\pi)(1-\alpha)$, by the first part of the theorem. By subtraction we have

$$\begin{aligned} L_{1,\alpha}(f,\mathcal{F}) &= \pi^{-1}(L_{X,\alpha}(f,\mathcal{F}) + (1-\pi)(R_0(f)-\alpha)) \\ &\leq \pi^{-1}(L_{X,\alpha}(f,\mathcal{F}) + (1-\pi)\varepsilon). \end{aligned}$$

■

4 Statistical performance guarantees

Theorem 1 suggests that we may estimate the solution to (1) by solving an "artificial" binary classification problem, treating $x_1,\ldots,x_m$ as one class and $x_{m+1},\ldots,x_{m+n}$ as the other. If a learning rule is consistent or achieves certain rates of convergence for the Neyman-Pearson classification problem $X \sim P_0$ vs. $X \sim P_X$ [Cannon et al., 2002, Scott and Nowak, 2005], then those properties will hold for the same learning rule viewed as a solution to $X \sim P_0$ vs. $X \sim P_1$. In other words, if $L_{X,\alpha}, \varepsilon \to 0$, then $L_{1,\alpha} \to 0$ at the same rate. Although π will not affect the rate of convergence, Theorem 1 suggests that small π makes the problem harder in practice, a difficulty which cannot be avoided.

As an illustrative example, we consider the case of a fixed set of classifiers $\mathcal{F}$ having finite VC-dimension [Vapnik, 1998] and consider

$$\begin{aligned} \widehat{f}_\tau = \arg\min_{f\in\mathcal{F}} \widehat{R}_X(f) \\ \text{s.t. } \widehat{R}_0(f) \leq \alpha + \tau, \end{aligned}$$

where $\widehat{R}$ is the empirical version of the corresponding error quantity. Define the precision of a classifier f for class i as $Q_i(f) = P_{XY}(Y=i|f(X)=i)$ (the higher the precision, the better the performance). Then we have the following result bounding the difference of the quantities R_i and Q_i to their optimal values over $\mathcal{F}$:

Theorem 2 *Assume the nominal and unlabeled data are i.i.d. realizations of their respective distributions, and that the two samples are independent of each other. Let $\mathcal{F}$ be a set of classifiers of VC-dimension V. Assume property* **(A)** *is satisfied and denote by f^* the optimal classifier in $\mathcal{F}$ with respect to the criterion in* (1). *Fixing $\delta > 0$, define $\varepsilon_k = \sqrt{\frac{V\log k - \log\delta}{k}}$. There exist absolute constants c, c' such that, if we choose $\tau = c\varepsilon_n$, the following bounds hold with probability $1-\delta$:*

$$R_0(\widehat{f}_\tau) - \alpha \leq c'\varepsilon_n\,; \tag{3}$$

$$R_1(\widehat{f}_\tau) - R_1(f^*) \leq c'\pi^{-1}(\varepsilon_n + \varepsilon_m) \tag{4}$$

$$Q_i(f^*) - Q_i(\widehat{f}_\tau) \leq \frac{c'}{P(f^*(X)=i)}(\varepsilon_n + \varepsilon_m)\,,\ i = 0,1\,. \tag{5}$$

The proof is given in Appendix A. The primary technical ingredients in the proof are Theorem 3 of Scott and Nowak [2005] and the learning reduction of Theorem 1 above. The above theorem shows that the procedure is consistent inside the class $\mathcal{F}$ for all criteria considered, i.e., these quantities decrease (resp. increase) asymptotically to their value at f^*. This is in contrast to the statistical learning bounds previously obtained [Liu et al., 2002, Thm. 2], which do not imply consistency.

Following Scott and Nowak [2005], by extending suitably the argument and the method in the spirit of structural risk minimization over a sequence of classes $\mathcal{F}_k$ having the universal approximation property, we can conclude that this method is universally consistent (i.e. relevant quantities converge to their value at f^*, where f^* is the solution of (1) over the set of all possible classifiers). Therefore, although technically simple, the reduction result of Theorem 1 allows us to deduce stronger results than the existing ones concerning this problem. This can be paralleled with the result that inductive novelty detection can be reduced to classification against uniform data [Steinwart et al., 2005], which made the statistical learning study of that problem significantly simpler.

It is interesting to note that the multiplicative constant in front of the rate of convergence of the precision criteria is $P_X(f^*(X) = i)^{-1}$ rather than π^{-1} for R_1. In particular $P_X(f^*(X) = 0) \geq (1-\pi)(1-\alpha)$, so that the convergence rate for class 0 precision is not significantly affected as $\pi \to 0$. Similarly $P_X(f^*(X) = 1) \geq (1-\pi)\alpha$, so the convergence rate for class 1 precision depends more crucially on the (known) α than on π.

Under an additional minor condition it is possible to show (the details are given at the end of Appendix B) that under the constraint $R_0(f) \leq \alpha$, the best attainable precision for class 0 in the set $\mathcal{F}$ is attained by $f = f^*$. Therefore, in (5) ($i = 0$), we are really comparing the precision of $\widehat{f_\tau}$ against the best possible class 0 precision given the FPR constraint. On the other hand, it does not make sense to consider the best attainable class 1 precision under an upper constraint on R_0, since we can have both $R_0 \to 0$ and $Q_1 \to 1$ by only rejecting a vanishingly small proportion of very sure novelties. But it can easily be seen that f^* realizes the best attainable class 1 precision under the *equality* constraint $R_0(f) = \alpha$.

We emphasize that the above result is but one of many possible theorems that could be deduced from the learning reduction; other results from Neyman-Pearson classification could also be applied. We also remark that, although the previous theorem corresponds to the semi-supervised setting, an analogous transductive result is easily obtained by incorporating an additional uniform deviation bound relating the empirical error rates on the unlabeled data to the true error rates.

5 The case $\pi = 0$ and a hybrid method

The preceding analysis only applies when $\pi > 0$. When $\pi = 0$, the learning reduction is trying to classify between two identical distributions, and the resulting decision rule could be arbitrarily poor. In this situation, perhaps the best we can expect is to perform as well as an inductive method. Therefore we ask the following question: Can we devise a method which, having no knowledge of π, shares the properties of the learning reduction above when $\pi > 0$, and reduces to the inductive approach otherwise? Our answer to the question is “yes” under fairly general conditions.

The intuition behind our approach is the following: As noted earlier, the inductive approach essentially performs density level set estimation. Therefore, level set estimation can be achieved by generating an artificial uniform sample and performing weighted binary classification against the nominal data. Thus, our approach is to sprinkle a vanishingly small proportion of uniformly distributed data among the unlabeled data. When $\pi = 0$, the uniform points will influence the final decision rule, but when $\pi > 0$, they will be swamped by the actual novelties.

To formalize this approach, let $0 < p_n < 1$ be a sequence tending to zero. Assume that S is a set which is known to contain the support of P_0 (obtained, e.g., through support estimation), and let P_2 be the uniform distribution on S. Consider the following procedure: Let $k \sim \text{binom}(n, p_n)$. Draw k independent realizations from P_2, and redefine $x_{m+1}, \ldots, x_{m+k}$ to be these values. (In practice, the uniform data would simply be appended to the unlabeled data, so that information is not erased. The present procedure, however, is slightly simpler to analyze.)

The idea now is to apply the SSND learning reduction from before to this modified unlabeled data. Toward this end, we introduce the following notations. For simplicity, we do not explicitly indicate the underlying class $\mathcal{F}$. We refer to any data point that was drawn from either P_1 or P_2 as an *operative* novelty. The proportion of operative novelties in the modified unlabeled sample is $\tilde{\pi} := \pi(1 - p_n) + p_n$. The distribution of operative novelties is $\tilde{P}_1 := \frac{\pi(1-p_n)}{\tilde{\pi}} P_1 + \frac{p_n}{\tilde{\pi}} P_2$, and the overall distribution of the modified unlabeled data is $\tilde{P}_X := \tilde{\pi}\tilde{P}_1 + (1 - \tilde{\pi})P_0$. Let $R_2, R^*_{2,\alpha}, \tilde{R}_1, \tilde{R}^*_{1,\alpha}, \tilde{R}_X$, and $\tilde{R}^*_{X,\alpha}$ be defined in terms of $P_2, \tilde{P}_1$, and $\tilde{P}_X$, respectively, in analogy to the definitions in Section 3. Also denote $L_{2,\alpha}(f) = R_2(f) - R^*_{2,\alpha}$, $\tilde{L}_{1,\alpha}(f) = \tilde{R}_1(f) - \tilde{R}^*_{1,\alpha}$, and $\tilde{L}_{X,\alpha} = \tilde{R}_X(f) - \tilde{R}^*_{X,\alpha}$.

By applying Theorem 1 to the modified data, we immediately conclude that if $R_0(f) \leq \alpha + \varepsilon$, then

$$\tilde{L}_{1,\alpha}(f) \leq \frac{1}{\tilde{\pi}}(\tilde{L}_{X,\alpha}(f) + (1 - \tilde{\pi})\varepsilon) = \frac{1}{\tilde{\pi}}(\tilde{L}_{X,\alpha}(f) + (1 - \pi)(1 - p_n)\varepsilon). \tag{6}$$

By previously cited results on Neyman-Pearson classification, the quantities on the right-hand side can be made arbitrarily small as m and n grow. The following result translates this bound to the kind of guarantee we are seeking.

Theorem 3 *Assume* **(A)** *holds. Let f be a classifier with $R_0(f) \leq \alpha + \varepsilon$. If $\pi = 0$, then*

$$L_{2,\alpha}(f) \leq p_n^{-1}(\tilde{L}_{X,\alpha}(f) + (1 - p_n)\varepsilon).$$

If $\pi > 0$, then

$$L_{1,\alpha}(f) \leq \frac{1}{\pi(1 - p_n)}(\tilde{L}_{X,\alpha}(f) + (1 - \pi)(1 - p_n)\varepsilon + p_n).$$

To interpret the first statement, note that $L_{2,\alpha}(f)$ is the inductive regret. The bound implies that $L_{2,\alpha}(f) \to 0$ as long as both $\varepsilon = R_0(f) - \alpha$ and $\tilde{L}_{X,\alpha}(f)$ tend to zero *faster than p_n*.

This suggests taking p_n to be a sequence tending to zero slowly. The second statement is similar to the earlier result in Theorem 1, but with additional factors of p_n. These factors suggest choosing p_n tending to zero rapidly, in contrast to the first statement, so in practice some balance should be struck.

Proof If $\pi = 0$, then $\tilde{L}_{1,\alpha} = L_{2,\alpha}$ and the first statement follows trivially from (6). To prove the second statement, denote $\beta_n := \frac{\pi(1-p_n)}{\tilde{\pi}}$, and observe that

$$\begin{aligned}
\tilde{R}^*_{1,\alpha} &= \inf_{R_0(f)\leq\alpha} \tilde{R}_1(f) \\
&= \inf_{R_0(f)\leq\alpha} [\beta_n R_1(f) + (1-\beta_n) R_2(f)] \\
&\leq \beta_n R^*_{1,\alpha} + (1-\beta_n).
\end{aligned}$$

Therefore

$$\begin{aligned}
\tilde{L}_{1,\alpha}(f) &= \tilde{R}_1(f) - \tilde{R}^*_{1,\alpha} \\
&\geq \beta_n R_1(f) + (1-\beta_n) R_2(f) - \beta_n R^*_{1,\alpha} - (1-\beta_n) \\
&\geq \beta_n (R_1(f) - R^*_{1,\alpha}) - (1-\beta_n) \\
&= \beta_n L_{1,\alpha}(f) + (1-\beta_n)
\end{aligned}$$

and we conclude

$$\begin{aligned}
L_{1,\alpha}(f) &\leq \frac{1}{\beta_n}\tilde{L}_{1,\alpha} + \frac{1-\beta_n}{\beta_n} \\
&\leq \frac{1}{\pi(1-p_n)}(\tilde{L}_{X,\alpha}(f) + (1-\pi)(1-p_n)\varepsilon + p_n).
\end{aligned}$$

■

We remark that this hybrid procedure could be applied with any prior distribution on novelties besides uniform. In addition, the hybrid approach could also be practically useful when n is small, assuming the artificial points are appended to the unlabeled sample.

6 Estimating π and testing for $\pi = 0$

We now turn to estimating the proportion π of novelties in the contaminated distribution P_X . For this purpose it should first be noted at this point that without additional assumptions, π is not an identifiable parameter in our model. To see this, consider the idealized case where we have an infinite amount of nominal and contaminated data, so that we have perfect knowledge of P_0 and P_X . Assuming the decomposition $P_X = (1-\pi)P_0 + \pi P_1$ holds, note that any alternate decomposition of the form $P_X = (1-\pi-\gamma)P_0 + (\pi+\gamma)P'_1$, with $P'_1 = (\pi+\gamma)^{-1}(\pi P_1 + \gamma P_0)$, and $\gamma \in [0, 1-\pi]$, is equally valid. Because the most important feature of the model is that we have no direct knowledge of P_1 , we cannot decide which representation is the "correct" one; we could not even exclude *a priori* the case where $\pi = 1$ and $P_1 = P_X$. The previous results established in Theorems 1-3 are valid for whatever underlying representation is assumed to be correct. For the estimation of the proportion of novelties however, it makes sense to define π as the *minimal* proportion of novelties that can explain the difference between P_0 and P_X. First we introduce the following definition:

Definition 4 *Assume P_0 P_1 are probability distribution functions on the space $\mathcal{X}$. We call P_1 a* proper novelty distribution *with respect to P_0 if there exists no decomposition of the form $P_1 = (1-\gamma)Q + \gamma P_0$ where Q is some probability distribution and $\gamma > 0$.*

This defines a proper novelty distribution P_1 as one that cannot be confounded with P_0 – it cannot be represented as a (nontrivial) mixture of P_0 with another distribution.

The next result establishes a canonical decomposition of the contaminated distribution into a mixture of nominal data and proper novelties. As a consequence the proportion π^* of proper novelties, and therefore the proper novelty distribution P_1 itself, are well-defined (i.e. identifiable) given the knowledge of the (arbitrary) nominal and contaminated distributions.

Proposition 5 *Assume P_0, P_X are probability distributions on the space $\mathcal{X}$. Then there is a unique $\pi^* \in [0,1]$ and P_1 such that the decomposition $P_X = (1-\pi^*)P_0 + \pi^* P_1$ holds, and that P_1 is a proper novelty distribution wrt. P_0; moreover*

$$\pi^* := \min\{\alpha \in [0,1] : \exists Q \textit{ probability distribution: } P_X = (1-\alpha)P_0 + \alpha Q\}\,.$$

The proof is given in Appendix A. From now on we assume that π and P_1 are the proportion and distribution of proper novelties of P_X with respect to P_0.

6.1 Population case

We now want to relate the estimation of π to quantities previously introduced and problem (1). We first treat the population case and optimal novelty detection over the set of all possible classifiers.

Theorem 6 *For any classifier f, we have the inequality*

$$\pi \geq 1 - \frac{R_X(f)}{1 - R_0(f)}\,. \tag{7}$$

Optimizing this bound over a set of classifiers $\mathcal{F}$ under the FPR constraint $R_0(f) \leq \alpha$ yields for any $\alpha > 0$:

$$\pi \geq 1 - \frac{R^*_{X,\alpha}(\mathcal{F})}{1-\alpha}\,.$$

Furthermore, if $\mathcal{F}$ is a set of possibly randomized classifiers containing all deterministic classifiers,

$$\pi = 1 - \inf_{\alpha \in [0,1)} \frac{R^*_{X,\alpha}(\mathcal{F})}{1-\alpha}\,. \tag{8}$$

Proof For the first part, just write for any classifier f

$$\begin{aligned} 1 - R_X(f) &= P_X(f(X)=1) \\ &= (1-\pi)P_0(f(X)=1) + \pi P_1(f(X)=1) \\ &\leq (1-\pi)R_0(f) + \pi , \end{aligned}$$

resulting in the first inequality in the theorem. Under the constraint $R_0(f) \leq \alpha$, this inequality then yields

$$\pi \geq 1 - \frac{R_X(f)}{1-R_0(f)} \geq 1 - \frac{R_X(f)}{1-\alpha};$$

optimizing the bound under the constraint yields the second inequality.

We establish in Lemma 13 in Appendix A that for any $\varepsilon > 0$ there exists a deterministic classifier f such that $R_0(f) < 1$ and $R_1(f)/(1-R_0(f)) \leq \varepsilon$. Put $\alpha = R_0(f)$; we then have

$$R^*_{X,\alpha}(\mathcal{F}) \leq R_X(f) = (1-\pi)(1-\alpha) + \pi R_1(f),$$

implying

$$\pi \geq 1 - \frac{R^*_{X,\alpha}(\mathcal{F})}{1-\alpha} \geq \pi\left(1 - \frac{R_1(f)}{1-R_0(f)}\right) \geq \pi(1-\varepsilon),$$

which establishes the last claim of the theorem. ■

6.2 Distribution-free lower confidence bounds on π

In the last part of Theorem 6, if we assume that the function $\alpha \mapsto R^*_{X,\alpha}(\mathcal{F})/(1-\alpha)$ is nonincreasing (a common regularity assumption; see Appendix B for a discussion of how this condition can always be ensured by considering possibly randomized classifiers), then $\alpha \mapsto R^*_{X,\alpha}(\mathcal{F})$ is left differentiable at $\alpha = 1$ and (8) reduces to

$$\pi = 1 - \left.\frac{dR^*_{X,\alpha}(\mathcal{F})}{d\alpha}\right|_{\alpha=1^-}.$$

This suggests estimating π by estimating the slope of $R^*_{X,\alpha}(\mathcal{F})$ at its right endpoint (where we recall that $\mathcal{F}$ must contain at least all deterministic classifiers). This can be related to the problem of estimating a monotone density at its right endpoint [Kulikov and Lopuhaä, 2006, Langaas et al., 2005]. Rather than pursue this approach here, however, we instead employ learning-theoretic techniques to use (7) for deriving a lower confidence bound on π:

Theorem 7 *Consider a classifier set $\mathcal{F}$ for which we assume a uniform error bound of the following form is available: for any distribution Q on $\mathcal{X}$, with probability at least $1-\delta$ over the draw of an i.i.d. sample of size n according to Q, we have*

$$\forall f \in \mathcal{F} \quad \left| Q(f(X)=1) - \widehat{Q}(f(X)=1) \right| \leq \varepsilon_n(\mathcal{F},\delta), \tag{9}$$

where $\widehat{Q}$ denotes the empirical distribution built on the sample.

Then the following quantity is a lower bound on π *with probability at least* $(1-\delta)^2 \geq 1-2\delta$ *(over the draw of the nominal and unlabeled samples) :*

$$\widehat{\pi}^-(\mathcal{F},\delta) = 1 - \inf_{f\in\mathcal{F}} \frac{\widehat{R}_X(f)+\varepsilon_n}{(1-\widehat{R}_0(f)-\varepsilon_m)_+}, \tag{10}$$

where the ratio is formally defined to be 1 *whenever the denominator is* 0 .

Note that if we define $\widehat{f}_\alpha = \arg\min_{f\in\mathcal{F}} \widehat{R}_X(f)$ under the constraint $\widehat{R}_0(f) \leq \alpha$, this can be rewritten

$$\widehat{\pi}^-(\mathcal{F},\delta) = 1 - \inf_{\alpha\in[0,1]} \frac{\widehat{R}_X(\widehat{f}_\alpha)+\varepsilon_n}{(1-\widehat{R}_0(\widehat{f}_\alpha)-\varepsilon_m)_+}.$$

There are two balancing forces at play here. From the population version, we know that we would like to have α as close as possible to 1 for estimating the derivative of $R^*_{X,\alpha}(\mathcal{F})$ at $\alpha = 1$. This is balanced by the estimation error which makes estimations close to $\alpha = 1$ unreliable because of the denominator. Taking the inf along the curve takes in a sense the best available tradeoff.

Proof As in the proof of the previous result, write for any classifier f:

$$P_X(f(X)=1) \leq (1-\pi)P_0(f(X)=1)+\pi,$$

from which we deduce after applying the uniform bound

$$\begin{aligned} 1-\widehat{R}_X(f)-\varepsilon_n &= \widehat{P}_X(f(X)=1)-\varepsilon_n \\ &\leq (1-\pi)(\widehat{R}_0(f)+\varepsilon_m)+\pi, \end{aligned}$$

which can be solved whenever $1-\widehat{R}_0(f)-\varepsilon_m > 0$. ■

The following result shows that $\widehat{\pi}^-(\mathcal{F},\delta)$, when suitably applied using a sequence of classifier sets $\mathcal{F}_1,\mathcal{F}_2,\ldots$ that have a universal approximation property leads to a strongly universally consistent estimate of the proportion π of proper novelties. The proof is given in Appendix A and relies on Theorem 7 in conjunction with the Borel-Cantelli lemma.

Theorem 8 *Consider a sequence* $\mathcal{F}_1,\mathcal{F}_2,\ldots$ *of classifier sets having the following universal approximation property: for any measurable function* $f^*:\mathcal{X}\to\{0,1\}$, *and any distribution* Q, *we have*

$$\liminf_{k\to\infty}\inf_{f\in\mathcal{F}_k} Q(f(X)\neq f^*(X)) = 0.$$

Suppose also that each class $\mathcal{F}_k$ *has finite VC-dimension* V_k, *so that for each* $\mathcal{F}_k$ *we have a uniform confidence bound of the form* (9) *for* $\varepsilon_n(\mathcal{F}_k,\delta) = 3\sqrt{\frac{V_k\log(n+1)-\log\delta/2}{n}}$. *Define*

$$\widehat{\pi}^-(\delta) = \sup_k \widehat{\pi}^-\left(\mathcal{F}_k,\delta k^{-2}\right).$$

If $\delta = (mn)^{-2}$, *then* $\widehat{\pi}^-$ *converges to* π *almost surely as* $m,n\to\infty$.

6.3 There are no distribution-free upper bounds on π

The lower confidence bounds $\widehat{\pi}^{-}(\mathcal{F}, \delta)$ and $\widehat{\pi}^{-}(\delta)$ are distribution-free in the sense that they hold regardless of P_0, P_1 and π. We now argue that distribution-free upper confidence bounds do not generally exist.

We define a *distribution-free upper confidence bound* $\widehat{\pi}^{+}(\delta)$ to be a function of the observed data such that, for any P_0, any proper novelty distribution P_1, and any $\pi < 1$, we have $\widehat{\pi}^{+}(\delta) \geq \pi$ with probability $1-\delta$ over the draw of the two samples.

We will show that such a universal upper bound does not exist unless it is trivial. The reason is that the novel distribution can be arbitrarily hard to distinguish from the nominal distribution. Looking at Section 6, this means that the slope of the straight line between $(\alpha, P_X(f^*_\alpha = 1))$ and $(1,1)$ can be made arbitrarily close to one for very small values of α while its derivative at $\alpha = 1$ remains bounded away from one. We can detect with some certainty that there is some proportion of novelties in the contaminated data (see Corollary 11 below), but we can never be sure that there are no novelties. This situation is similar to the philosophy of significance testing: one can never accept the null hypothesis, but only have insufficient evidence to reject it.

We will say that the nominal distribution P_0 is *weakly diffuse* if for any $\gamma > 0$ there exists a set A such that $0 < P_0(A) < \gamma$. We say an upper confidence bound $\widehat{\pi}^{+}(\delta)$ is *non-trivial* if there exists a weakly diffuse nominal distribution P_0, a novelty distribution P_1, constants $\pi > 0$, $\delta > 0$ such that

$$P(\widehat{\pi}^{+}(\delta) < 1) > \delta,$$

where the probability is over the joint draw of nominal and contaminated samples. This assumption demands that there is at least a specific setting where the upper bound $\widehat{\pi}^{+}(\delta)$ is significantly different from the trivial bound 1, meaning that it is bounded away from 1 with larger probability than its allowed probability of error δ.

Theorem 9 *There exists no distribution-free, non-trivial upper confidence bound on π.*

The proof appears in Appendix A. The non-triviality assumption is quite weak and relatively intuitive. The only not directly intuitive assumption is that P_0 should be weakly diffuse, which is satisfied for all distributions having a continuous part. This assumption effectively excludes finite state spaces. We believe it is possible to obtain a non-trivial upper confidence bound on π on a finite state space.

Corollary 10 *The rate of convergence of any distribution-free lower bound $\widetilde{\pi}^{-}$ towards π can be arbitrarily slow.*

Proof If there was a universally valid upper bound δ_n on the convergence rate of $\widetilde{\pi}^{-}$, then $\widetilde{\pi}^{-} + \delta_n$ would be a distribution-free upper confidence bound on π. ■

To achieve some prescribed rate of convergence, some assumptions on the generating distributions must be made. This parallels the estimation of the Bayes risk in classification [Devroye, 1982].

6.4 Testing for $\pi = 0$

The lower confidence bound on π can also be used as a test for $\pi = 0$, i.e., a test for whether there are any novelties in the test data:

Corollary 11 *Let $\mathcal{F}$ be a set of classifiers. If $\widehat{\pi}^{-}(\mathcal{F}, \delta) > 0$, then we may conclude, with confidence $1 - \delta$, that the unlabeled sample contains novelties.*

It is worth noting that testing this hypothesis is equivalent to testing if P_0 and P_X are the same distribution, which is the classical two-sample problem in an arbitrary input space. This problem has recently generated attention in the machine learning community [Gretton et al., 2007], and the approach proposed here, using arbitrary classifiers, seems to be new. Our confidence bound could of course also be used to test the more general hypothesis $\pi \leq \pi_0$ for a prescribed π_0, $0 \leq \pi_0 < 1$.

Note that, by definition of $\widehat{\pi}^{-}(\mathcal{F}, \delta)$, testing the hypothesis $\pi = 0$ using the above lower confidence bound for π is equivalent to searching the classifier space $\mathcal{F}$ for a classifier f such that the proportions of predictions of 0 and 1 by f differ on the two samples in a statistically significant manner. Namely, for a classifier f belonging to a class $\mathcal{F}$ for which we have a uniform bound of the form (9), we have the lower bound $P_X(f(X) = 1) \geq \widehat{P}_X(f(X) = 1) - \varepsilon_n$ and the upper bound $P_0(f(X) = 1) \leq \widehat{P}_0(f(X) = 1) + \varepsilon_m$ (both bounds valid simultaneously with probability at least $1 - \delta$). If the difference of the bounds is positive we conclude that we must have $P_X \neq P_0$, hence $\pi > 0$. This difference is precisely what appears in the numerator of $\widehat{\pi}^{-}(\mathcal{F}, \delta)$ in (10). Furthermore, if this numerator is positive then so is the denominator since it is always larger. In the end, $\widehat{\pi}^{-}(\mathcal{F}, \delta) > 0$ is equivalent to

$$\sup_{f \in \mathcal{F}} \Big((\widehat{P}_X(f(X) = 1) - \varepsilon_n) - (\widehat{P}_0(f(X) = 1) + \varepsilon_m) \Big) > 0 \, .$$

7 Relationship between SSND and multiple testing

In this section, we show how SSND offers powerful generalizations of the standard p-value approach to multiple testing under the widely used “random effects” model, as considered for example by Efron et al. [2001].

7.1 Multiple testing under the random effects model

In the multiple testing framework, a finite family $(H_1, \ldots, H_K)$ of null hypotheses to test is fixed; from the observation of some data X , a decision $D(H_i, X) \in \{0, 1\}$ must be taken for each hypothesis, namely whether (given the data) hypothesis H_i is deemed to be false ($D(H_i, X) = 1$, hypothesis rejected) or true ($D(H_i, X) = 0$, hypothesis not rejected). A typical exemplary application domain is that of microarray data analysis, where each null hypothesis H_i corresponds to the absence of a difference in expression levels of gene i

in a comparison between two experimental situations. A rejected null hypothesis then indicates such a differential expression for a specific gene, and is called a *discovery* (since differentially expressed genes are those of interest). However, the number of null hypotheses to test is very large, for example $K \simeq 4.10^4$ in the gene expression analysis, and the probability of rejecting by chance a null hypothesis must be strictly controlled.

In the standard setting for multiple testing, it is assumed that a testing statistic $Z_i(X) \in \mathbb{R}$ has been fixed for each null hypothesis H_i, and that its marginal distribution is known when H_i is true. This statistic can then be normalized (e.g. by monotone transform) to take the form of a *p-value*. A p-value is a function $p_i(X)$ of the data such that, if the corresponding null hypothesis H_i is true, then $p_i(X)$ has a uniform marginal distribution on $[0,1]$. In this setting, it is expected that the rejection decisions $D(H_i,X)$ are taken based on the observed p-values $(p_1,\ldots,p_K)$ rather than on the raw data. In fact, in most cases it is assumed that the decisions take the form $D(H_i,X) = \mathbf{1}_{p_i \leq \widehat{T}}$, where $\widehat{T}$ is a data-dependent threshold. Further, simplifying distributional assumptions on the family of p-values are often posited. A common distribution model called *random effects* assumes that the veracity of hypothesis H_i is governed by an underlying latent variable h_i and reads as follows:

- the variables $h_i \in \{0,1\}$, $1 \leq i \leq K$ are i.i.d. Bernoulli with parameter π
- the variables p_i are independent, and conditionally to $(h_1,\ldots,h_K)$ have distribution

$$p_i \sim \begin{cases} \text{Uniform}[0,1]\,, & \text{if } h_i = 0 \\ P_1\,, & \text{if } h_i = 1\,. \end{cases}$$

Under the random effects model, the p-values thus follow a mixture distribution $(1-\pi)U[0,1]+\pi P_1$ on the interval $[0,1]$ and can be seen as a contaminated sample, while the variables h_i play the role of the unknown labels. It should now be clear that the above model is in fact a *specification* of the SSND model, with the following additional assumptions:

1. The observation space is the interval $[0,1]$;
2. The nominal distribution P_0 is known to be exactly uniform on $[0,1]$ (equivalently, the nominal distribution is uniform and the nominal sample has infinite size)
3. The class of novelty detectors considered is the set of intervals of the form $[0,t], t \in [0,1]$.

Therefore, the results developed in this paper can apply to the more restricted setting of multiple testing under the random effects model as well. In particular, the estimator $\widehat{\pi}^-(\mathcal{F},\delta)$ developed in Section 6, when specified under the above additional conditions, recovers the methodology of non-asymptotic estimation of $1-\pi$ which was developed by Genovese and Wasserman [2004], Section 3, and our notion of proper novelty distribution recovers their notion of *purity* in that setting (and has somewhat more generality, since they assumed P_1 to have a density).

There are several interesting benefits in considering for the purpose of multiple testing the more general SSND model developed here. First, it can be unrealistic in practice to assume that the distribution of the p-values is known exactly under each one of the null hypotheses: instead, only assuming the knowledge of a reference sample under controlled experimental conditions as in the SSND model is often more realistic. Secondly, the restriction to decision sets of the form $\{p_i \leq t\}$ can also be questionable. For a single test, decision regions of this form are optimal (in the Neyman-Pearson sense) only if the likelihood ratio of the alternative to the null is decreasing, which amounts to assuming that the alternative distribution P_1 has a decreasing density.

This assumption has been criticized in some recent work. A simple example of a situation where this assumption fails is in the framework of z or t-tests, i.e., the null distribution of the statistic (before rescaling into p-values) is a standard Gaussian or a Student t-distribution, and the corresponding p-value function is the usual one- or two-sided p-value. If the alternative distribution P_1 is a mixture of Gaussians (resp. of noncentral t distributions), optimal rejection regions for the original statistic are in general a finite union of disjoint intervals and do not correspond to level sets of the p-values. In order to counter this type of problem, Sun and Cai [2007] suggest to estimate from the data the alternate density and the proportion of true null hypotheses, and use these estimates directly in a plug-in likelihood ratio based test. Chi [2007] develops a procedure based on growing rejection intervals around a finite number of fixed control points in $[0,1]$. In both cases, an asymptotic theory is developed. Both of these procedures are more flexible than using only rejection intervals of the form $[0,t]$ and aim at adaptivity with respect to the alternative distribution P_1. Finally, the remaining restriction that effective observations (the p-values) belong to the unit interval was also put into question by Chi [2008], who considered a setting of multidimensional p-values belonging to $[0,1]^d$. The distribution was still assumed to be uniform under the corresponding null hypothesis, although this seems an even less realistic assumption than in dimension one: in this framework, the use of a reference "nominal" sample under the null distribution seems even more relevant.

The framework developed in the present paper allows to cover at once these different types of extensions rather naturally by just considering a richer class $\mathcal{F}$ of candidate classifiers (or equivalently in this setting, rejection regions), and provides a non-asymptotical analysis of their behavior using classical learning theoretical tools such as VC inequalities. Furthermore, such non-asymptotic inequalities can also give rise to adaptive and consistent model selection for the set of classifiers using the structural risk minimization principle, a topic that was not addressed previously for the extensions mentioned above.

7.2 SSND with controlled FDR

One remaining important difference between the SSND setting studied here and that of multiple testing is that our main optimization problem (1) is under a false positive rate constraint $R_0(f) \leq \alpha$, while most recent work on multiple testing generally imposes a constraint on the false discovery rate (FDR) instead. If we denote $\mathrm{Pos}(f) = \widehat{P}_X(f(X)=1)$ and $\mathrm{FP}(f) = \widehat{P}_{XY}(f(X)=1, Y=0)$ the proportion of reported novelties and the proportion of false discoveries on the contaminated sample, respectively, then the false discovery proportion (FDP) is defined as $\mathrm{FDP}(f) = \mathrm{Pos}(f)/\mathrm{FP}(f)$ (taken to be zero if the denominator vanishes), and the FDR is defined as $\mathrm{FDR}(f) = E[\mathrm{FDP}(f)]$. Some classical variations of this quantity are the positive FDR, $\mathrm{pFDR}(f) = E[\mathrm{FDP}(f)|\mathrm{Pos}(f) > 0]$ and the marginal FDR, $\mathrm{mFDR}(f) = E[\mathrm{FP}(f)]/E[\mathrm{Pos}(f)]$. Under the mixture contamination model, it can be checked that $\mathrm{pFDR}(f) = \mathrm{mFDR}(f) = P_{XY}(Y = 0|f(X) = 1)$ [Storey, 2003], hence also equal to one minus the precision for class 1 (as defined earlier in Section 4). The following result states explicit empirical bounds on these quantities:

Proposition 12 *Consider a classifier set $\mathcal{F}$ for which we assume uniform error bound of the following form is available: for any distribution Q on $\mathcal{X} \times \{0,1\}$, with probability at least $1-\delta$ over the draw of an i.i.d. sample of size n according to Q, both*

$$\forall f \in \mathcal{F} \quad \left|Q(f(X) = 1) - \widehat{Q}(f(X) = 1)\right| \leq \varepsilon_n(\mathcal{F},\delta), \tag{11}$$

and

$$\forall f \in \mathcal{F} \quad \left|Q(f(X) = 1, Y = 0) - \widehat{Q}(f(X) = 1, Y = 0)\right| \leq \varepsilon_n(\mathcal{F},\delta), \tag{12}$$

hold, where $\widehat{Q}$ denotes the empirical distribution built on the sample.

Then the following inequalities hold with probability at least $(1-\delta)^2 \geq 1-2\delta$ (over the draw of the nominal and unlabeled samples) :

$$\forall f \in \mathcal{F} \;\; \mathrm{mFDR}(f) = P_X(Y = 0|X = 1) \leq \frac{(\widehat{R}_0(f) + \varepsilon_m)(1 - \widehat{\pi}^-(\mathcal{F},\delta))}{(1 - \widehat{R}_X(f) - \varepsilon_n)_+},$$

and

$$\forall f \in \mathcal{F} \;\; \mathrm{FDP}(f) \leq \frac{(\widehat{R}_0(f) + \varepsilon_m)(1 - \widehat{\pi}^-(\mathcal{F},\delta)) + \varepsilon_n}{(1 - \widehat{R}_X(f))},$$

where $\widehat{\pi}^-(\mathcal{F},\delta)$ is defined in (10).

Note that equations (11) (12) holds as before with $\varepsilon_n(\mathcal{F},\delta) = c\sqrt{\frac{V \log n - \log\delta}{n}}$ when $\mathcal{F}$ has VC dimension V. In the interest of simplicity, we use the same bound ε_n for both uniform error assumptions. Separate bounds could also be adopted, allowing (11) to be slightly tighter. We also remark that since FDP is an empirical quantity based on the contaminated sample, the second bound is in fact a *transductive* bound rather than semi-supervised.

Proof The mFDR can be rewritten as $\mathrm{mFDR}(f) = P_0(f(X) = 1|Y = 0)P_{XY}(Y = 0)/P_X(f(X) = 1) = R_0(f)(1-\pi)/(1-R_X(f))$; in this expression we can plug in the lower bound for π of Theorem 7 and uniform bounds for $R_0(f)$ and $R_X(f)$ coming from assumption (11). The FDP can be written as $\mathrm{FDP}(f) = \widehat{P}_{XY}(f(X) = 1, Y = 0)/(1-\widehat{R}_X(f))$. Using assumption (12), the numerator can be upper bounded by $P_{XY}(f(X) = 1, Y = 0) + \varepsilon_n = R_0(f)(1-\pi) + \varepsilon_n$, and we can then use the same reasoning as for the first part. ■

Similarly to what was proposed in Section 4 under the false positive rate constraint, we can in this context consider to maximize $\widehat{R}_X(f)$ over $f \in \mathcal{F}$ subject to the constraint that the above empirical bound on the mFDR or FDP is less than α. This can then be suitably extended to a sequence of classes $\mathcal{F}_k$. While a full study of the resulting procedure is out of the scope of the present paper, we want to point out the important difference that the mFDR is necessarily lower bounded by $\inf_{x \in \mathcal{X}} P_{XY}(Y = 0|X = x)$ which is generally strictly positive. Hence, the required constraint may not be realizable if α is smaller than this lower bound, in which case the empirical procedure should return a failure statement with probability one as $n \to \infty$. A similar approach was also introduced recently by Scott et al. [2009], but under the stronger requirement that labeled examples from *both* classes are available.

Practical example. As a typical setting for multiple testing, assume that the observation space $\mathcal{X}$ is the real line, and that the reference distribution P_0 is exactly known to be a standard Gaussian. Then probability of events under P_0 can be computed exactly, and we can take $\varepsilon_m = 0$ in the previous bounds. Consider the set $\mathcal{F}_k$ of rejection regions (classifiers) made of a disjoint union of at most k intervals. Then $\mathcal{F}_k$ has VC dimension $2k$. To compute the relevant quantities such as the bound $\widehat{\pi}^-(\mathcal{F}_k, \delta)$, and the above bounds on the mFDR or the FDP, it is sufficient to consider intervals with endpoints belonging to $\{X_1, \ldots, X_n\}$. Furthermore, for all calculations it is sufficient to know, for each possible fixed value of $\widehat{R}_X$, what is the classifier achieving the lowest value of R_0. This can be accomplished using dynamic programming in time $O(kn^3)$.

8 Experiments

Despite previous work on learning with positive and unlabeled examples (LPUE), as discussed in Section 2, the efficacy of our proposed learning reduction has not been empirically demonstrated. To assess the impact of unlabeled data on novelty detection, we applied our framework to some datasets which are common benchmarks for binary classification. The first 13 datasets [Müller et al., 2001] are from `http://ida.first.fhg.de/projects/bench/` and the last five datasets [Chang and Lin, 2001] are from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Each dataset consists of both positive and negative examples. Furthermore, each dataset is replicated 100 times (except for image and splice, which are replicated 20 times), with each copy corresponding to a different random partitioning into training and test examples. All numerical results for a dataset were obtained by averaging across all partitions. The negative examples from the training set were taken to form the nominal sample, and the positive training examples were not used at all in the experiments. The datasets are summarized in Table 1. Here N_{train} and N_{test} are the sizes[1] of the training and test sets, respectively, while π_{base} is the proportion of positive examples in the combined training and test data. Thus, the average (across permutations) nominal sample size m is $(1 - \pi_{\text{base}})N_{\text{train}}$.

[1] The web and adult datasets were subsampled owing to their large size.

Data set	dim	N_{train}	N_{test}	π_{base}
banana	2	400	4900	0.45
breast-cancer	9	200	77	0.29
diabetes	8	468	300	0.35
flare-solar	9	666	400	0.55
german	20	700	300	0.30
heart	13	170	100	0.44
ringnorm	20	400	7000	0.50
thyroid	5	140	75	0.30
titanic	3	150	2051	0.32
twonorm	20	400	7000	0.50
waveform	21	400	4600	0.33
image	18	1300	1010	0.57
splice	60	1000	2175	0.48
ionosphere	34	251	100	0.64
mushrooms	112	4124	4000	0.48
sonar	60	108	100	0.47
adult	123	3000	3000	0.24
web	300	3000	3000	0.03

Table 1: Description of data sets. dim is the number of features, and N_{train} and N_{test} are the numbers of training and test examples. π_{base} is the proportion of positive examples (novelties) in the combined training and test data. Thus, the average (across permutations) nominal sample size m is $(1-\pi_{\text{base}})N_{\text{train}}$.

8.1 Experimental setup

We evaluated our methodology in two learning paradigms, comparing five learning methods across several values of π. The two learning paradigms are semi-supervised and transductive. For semi-supervised learning, the test data were divided into two halves. The first half was used as the contaminated, unlabeled data. The second half was used as an independent sample of contaminated data, not used in the learning stage, but only for independent evaluation of classifiers returned by each method. In particular, the second half of the test data was used to estimate the area under the ROC (AUC) of each method. Here, the ROC is the one which views P_0 as the null distribution and P_1 as the alternative. For transductive learning, the entire test set was treated as the unlabeled data, and was also used for evaluating the AUC.

The learning methods are the inductive approach, our proposed learning reduction, and three versions of the hybrid approach. The three hybrids correspond to $p_n = 1.0, 0.5, 0.1$, in which a uniform sample of size $100p_n$% of the unlabeled sample size is *appended* to the unlabeled data. We emphasize that each algorithm was implemented in the same way in the two learning paradigms; the only differences are the size of the contaminated sample, and how they are evaluated.

We implemented the inductive novelty detector using a thresholded kernel density estimate (KDE) with Gaussian kernel, and SSND using a plug-in KDE classifier. (To alleviate concerns that our inductive implementation is inadequate, we also tested the one-class support vector machine [Schölkopf et al., 2001] in several experimental settings, and found its performance to be very similar.) For each class, a single kernel bandwidth parameter was employed, and optimized by maximizing a cross-validation estimate of the AUC. This ROC is different from the one used to evaluate the methods (see above). In particular, it still views P_0 as the null distribution, but now the alternative distribution is taken to be the uniform distribution P_2 for the inductive detector (see Section 5; effectively we use a uniform random sample of size n in place of the unlabeled data), P_X for SSND, and the appropriate $\tilde{P}_X$ for the hybrid methods (see Section 5). Thus, the test label information was not used at any stage (prior to validation) by any of the methods.

We also compared the learning methods for several values of π. For semi-supervised learning, we examined $\pi = 0.5, \pi = 0.2, \pi = 0.1$, and $\pi = 0.0$. For transductive learning, we examined $\pi = 0.5, \pi = 0.2$, and $\pi = 0.1$. The case $\pi = 0.0$ cannot be evaluated in the transductive paradigm because there are no positive examples in the unlabeled data. For each value of π, we discarded just enough examples (either negative or positive) so that the desired proportion was achieved in the contaminated data. Note that the number of positive examples (novelties) in the contaminated sample could be very small. For the smallest datasets, in the semi-supervised setting and when $\pi = 0.1$, this number is less than 10.

8.2 Statistical summaries and methodology

The complete results are summarized in Tables 2 through 5. Tables 2 and 3 show the average AUC for each dataset and experimental setting, for the semi-supervised and transductive paradigms respectively. The inductive method is labeled Ind. Our learning reduction is labeled SSND or TND depending on the setting. The hybrid methods are labeled H(p) in Tables 2-3, and Hybrid(p) in Tables 4-5.

We followed the methodology of Demšar [2006] for comparing algorithms across multiple datasets. For each dataset and each experimental setting, the algorithms were ranked 1 (best) through 5 (worst) based on AUC. The Friedman test was used to determine, for each experimental setting, whether there was a significant difference in the average ranks of the five algorithms across the datasets. The average ranks and p-values are reported in Tables 4 and 5. The results indicate that there is a significant difference among the algorithms at the 0.1 significance level for all settings, with the exception of the transductive setting when $\pi = 0.1$.

When the Friedman test resulted in significant differences, we then performed a post-hoc Nemenyi test to assess when there was a significant difference between individual algorithms. For a five algorithm experiment on 18 datasets, with a significance level of 0.1, the critical difference for the Nemenyi test is 1.30. That is, when the average ranks of two algorithms differ by more than 1.30, their performance is deemed to be significantly different.

dataset	$\pi = 0.5$					$\pi = 0.2$				
	Ind.	SSND	H(1.0)	H(0.5)	H(0.1)	Ind.	SSND	H(1.0)	H(0.5)	H(0.1)
banana	0.924	0.939	0.931	0.933	0.936	0.924	0.915	0.924	0.923	0.921
breast-cancer	0.654	0.643	0.675	0.669	0.667	0.654	0.557	0.657	0.648	0.621
diabetes	0.744	0.782	0.770	0.772	0.776	0.744	0.684	0.724	0.727	0.717
flare-solar	0.674	0.661	0.664	0.660	0.662	0.674	0.629	0.641	0.643	0.642
german	0.628	0.703	0.693	0.696	0.704	0.628	0.582	0.633	0.632	0.636
heart	0.793	0.854	0.845	0.853	0.851	0.793	0.690	0.805	0.789	0.745
ringnorm	0.999	0.997	0.996	0.996	0.996	0.999	0.992	0.990	0.991	0.983
thyroid	0.985	0.966	0.964	0.967	0.955	0.985	0.889	0.929	0.940	0.943
titanic	0.628	0.643	0.636	0.644	0.643	0.628	0.612	0.636	0.634	0.628
twonorm	0.915	0.993	0.989	0.989	0.990	0.915	0.940	0.961	0.958	0.953
waveform	0.761	0.958	0.952	0.945	0.956	0.761	0.839	0.848	0.896	0.901
image	0.818	0.939	0.929	0.935	0.939	0.818	0.892	0.874	0.879	0.875
splice	0.415	0.935	0.905	0.921	0.932	0.415	0.702	0.613	0.764	0.785
ionosphere	0.256	0.926	0.839	0.921	0.922	0.256	0.695	0.475	0.607	0.704
mushrooms	0.945	1.000	1.000	1.000	1.000	0.945	0.999	0.999	0.999	0.999
sonar	0.688	0.752	0.757	0.764	0.764	0.688	0.595	0.682	0.683	0.646
adult	0.605	0.872	0.872	0.864	0.835	0.605	0.705	0.720	0.829	0.720
web	0.462	0.778	0.749	0.697	0.788	0.462	0.616	0.631	0.585	0.674

dataset	$\pi = 0.1$					$\pi = 0.0$				
	Ind.	SSND	H(1.0)	H(0.5)	H(0.1)	Ind.	SSND	H(1.0)	H(0.5)	H(0.1)
banana	0.924	0.891	0.922	0.919	0.913	0.924	0.540	0.919	0.905	0.785
breast-cancer	0.654	0.515	0.643	0.633	0.575	0.654	0.556	0.640	0.628	0.568
diabetes	0.744	0.605	0.699	0.700	0.692	0.744	0.494	0.689	0.669	0.657
flare-solar	0.674	0.571	0.624	0.629	0.626	0.674	0.471	0.613	0.603	0.611
german	0.628	0.548	0.623	0.624	0.602	0.628	0.522	0.595	0.608	0.592
heart	0.793	0.593	0.778	0.776	0.688	0.793	0.506	0.759	0.750	0.620
ringnorm	0.999	0.984	0.981	0.986	0.991	0.999	0.478	0.958	0.978	0.985
thyroid	0.985	0.786	0.884	0.906	0.895	0.985	0.590	0.852	0.869	0.795
titanic	0.628	0.591	0.632	0.634	0.621	0.628	0.443	0.630	0.628	0.572
twonorm	0.915	0.931	0.945	0.934	0.923	0.915	0.480	0.894	0.879	0.860
waveform	0.761	0.801	0.815	0.822	0.806	0.761	0.487	0.736	0.727	0.705
image	0.818	0.769	0.824	0.836	0.851	0.818	0.431	0.634	0.696	0.780
splice	0.415	0.630	0.518	0.584	0.625	0.415	0.523	0.447	0.493	0.493
ionosphere	0.256	0.618	0.438	0.488	0.575	0.256	0.520	0.392	0.431	0.486
mushrooms	0.945	0.995	0.992	0.998	0.996	0.945	0.566	0.972	0.980	0.982
sonar	0.688	0.556	0.658	0.652	0.615	0.688	0.510	0.628	0.643	0.587
adult	0.605	0.627	0.659	0.666	0.626	0.605	0.505	0.558	0.556	0.572
web	0.462	0.554	0.584	0.544	0.611	0.462	0.557	0.553	0.523	0.564

Table 2: AUC values for five novelty detection algorithms in the semi-supervised setting. ’H’ indicates a hybrid method.

dataset	$\pi = 0.5$					$\pi = 0.2$				
	Ind.	TND	H(1.0)	H(0.5)	H(0.1)	Ind.	TND	H(1.0)	H(0.5)	H(0.1)
banana	0.924	0.938	0.931	0.932	0.935	0.924	0.915	0.923	0.923	0.919
breast-cancer	0.663	0.673	0.662	0.662	0.670	0.663	0.615	0.649	0.659	0.630
diabetes	0.742	0.784	0.776	0.779	0.788	0.742	0.708	0.728	0.725	0.727
flare-solar	0.673	0.686	0.683	0.684	0.684	0.673	0.661	0.658	0.662	0.666
german	0.633	0.739	0.709	0.711	0.714	0.633	0.617	0.632	0.637	0.636
heart	0.796	0.869	0.856	0.856	0.864	0.796	0.716	0.811	0.794	0.788
ringnorm	0.999	0.997	0.996	0.996	0.996	0.999	0.993	0.989	0.991	0.983
thyroid	0.984	0.976	0.978	0.979	0.974	0.984	0.957	0.962	0.955	0.962
titanic	0.629	0.667	0.646	0.658	0.661	0.629	0.642	0.641	0.658	0.645
twonorm	0.915	0.993	0.990	0.990	0.990	0.915	0.940	0.961	0.961	0.956
waveform	0.771	0.960	0.953	0.947	0.957	0.771	0.847	0.850	0.900	0.905
image	0.845	0.955	0.949	0.949	0.953	0.845	0.897	0.889	0.891	0.901
splice	0.416	0.941	0.913	0.930	0.939	0.416	0.716	0.623	0.769	0.820
ionosphere	0.254	0.953	0.844	0.931	0.952	0.254	0.714	0.413	0.633	0.746
mushrooms	0.945	1.000	1.000	1.000	1.000	0.945	0.999	0.999	0.999	0.999
sonar	0.683	0.757	0.767	0.778	0.781	0.683	0.615	0.678	0.683	0.662
adult	0.606	0.875	0.873	0.865	0.835	0.606	0.687	0.736	0.847	0.739
web	0.464	0.810	0.758	0.727	0.788	0.464	0.644	0.639	0.590	0.667

dataset	$\pi = 0.1$				
	Ind.	TND	H(1.0)	H(0.5)	H(0.1)
banana	0.924	0.896	0.921	0.920	0.910
breast-cancer	0.663	0.564	0.687	0.642	0.598
diabetes	0.742	0.658	0.720	0.709	0.693
flare-solar	0.673	0.615	0.655	0.643	0.659
german	0.633	0.556	0.615	0.616	0.615
heart	0.796	0.626	0.792	0.784	0.729
ringnorm	0.999	0.985	0.973	0.986	0.992
thyroid	0.984	0.910	0.970	0.955	0.932
titanic	0.629	0.603	0.643	0.642	0.626
twonorm	0.915	0.933	0.943	0.937	0.923
waveform	0.771	0.813	0.821	0.823	0.808
image	0.845	0.888	0.870	0.871	0.880
splice	0.416	0.630	0.554	0.553	0.640
ionosphere	0.254	0.589	0.349	0.443	0.552
mushrooms	0.945	0.996	0.994	0.997	0.997
sonar	0.683	0.514	0.646	0.655	0.592
adult	0.606	0.658	0.681	0.684	0.629
web	0.464	0.567	0.573	0.538	0.604

Table 3: AUC values for five novelty detection algorithms in the transductive setting. ’H’ indicates a hybrid method.

π	Inductive	SSND	Hybrid(1.0)	Hybrid(0.5)	Hybrid(0.1)	p-value
0.0	1.89	4.39	2.72	2.89	3.11	0.000
0.1	2.83	4.00	2.83	2.28	3.06	0.023
0.2	3.28	3.83	2.61	2.56	2.72	0.071
0.5	4.28	1.94	3.44	3.00	2.33	0.000

Table 4: The comparison of average ranks of the five algorithms in the semi-supervised setting, by the Friedman test. The critical difference of the post-hoc Nemenyi test is 1.30 at a confidence level $\alpha = 0.1$.

π	Inductive	TND	Hybrid(1.0)	Hybrid(0.5)	Hybrid(0.1)	p-value
0.1	2.94	3.78	2.56	2.67	3.06	0.157
0.2	3.17	3.78	3.06	2.50	2.50	0.085
0.5	4.44	1.44	3.56	3.17	2.39	0.000

Table 5: The comparison of average ranks of the five algorithms in the transductive setting, by the Friedman test. The critical difference of the post-hoc Nemenyi test is 1.30 at a confidence level $\alpha = 0.1$.

8.3 Analysis of results

From the results presented in Tables 2-5, we draw the following conclusions.

1. The average ranks in Tables 4-5 conform to our expectations in many respects. SSND/TND outrank the inductive approach when $\pi = 0.5$, and inductive outranks semi-supervised when $\pi = 0.0$. At the intermediate values $\pi = 0.1$ and 0.2, hybrid methods achieve the best ranking.

2. The average ranks also reveal that the performance of the hybrid methods vary according to the value of π. As π increases, the best performing hybrid has a correspondingly smaller amount of auxiliary uniform data appended to the unlabeled sample. This also conforms to our expectations.

3. All tables indicate that the proposed methodology performs better in the transductive setting than the semi-supervised setting. A likely reason is that, in our experimental setup, TND sees twice as much unlabeled data as SSND.

4. When $\pi = 0.0$ in the semi-supervised experiments, SSND typically has an AUC around 0.5, which corresponds to random guessing. This makes sense, because it is essentially trying to classify between two realizations of the nominal distribution. From Tables 2 and 4 we see that the hybrid methods clearly improve upon SSND when $\pi = 0.0$.

5. For some datasets (splice, ionosphere, web), the inductive method does worse than random guessing, but our methods do not. In each case, our methods yield dramatic increases in AUC.

6. The benefits of unlabeled data increase with dimension. In particular, SSND and TND tend to perform much better relative to the inductive approach on datasets of

dimension at least 18. This is especially evident in the second half of the datasets, which even show significant gains for $\pi = 0.1$. This trend suggests that as dimension increases, the assumption implicit in the inductive approach (that novelties are uniform where they overlap the support of the nominal distribution) breaks down.

Figure 1 depicts a sampling of results comparing the inductive and semi-supervised methods, and highlights the impact of dimension. The top graph shows ROCs for a two-dimensional dataset where the two classes are fairly well separated, meaning the novelties lie in the tails of the nominal class, and $\pi = 0.5$. Not surprisingly, the inductive method is close to the semi-supervised method. The middle graph represents the 60-dimensional splice dataset, where the inductive method does worse than random guessing, yet SSND does quite well. The bottom graph in Figure 1 shows the results for the 21-dimensional waveform data for $\pi = 0.1$. Here the assumptions of the inductive approach are also evidently violated to some degree.

9 Conclusions

We have shown that semi-supervised novelty detection reduces to Neyman-Pearson classification. This allows us to leverage known performance guarantees for NP classification algorithms, and to import practical algorithms. We have applied techniques from statistical learning theory, such as uniform deviation inequalities, to establish distribution free performance guarantees for SSND, as well as a lower bound and consistent estimator for π, and test for $\pi = 0$. Our approach optimally adapts to the unknown novelty distribution, unlike inductive approaches, which operate as if novelties are uniformly distributed. We also introduced a hybrid method that has the properties of SSND when $\pi > 0$, and effectively reverts to the inductive method when $\pi = 0$.

Our analysis strongly suggests that in novelty detection, unlike traditional binary classification, unlabeled data are essential for attaining optimal performance in terms of tight bounds, consistency, and rates of convergence. In an extensive experimental study, we found that the advantages of our approach are most pronounced for high dimensional data. Our analysis and experiments confirm some challenges that seem to be intrinsic to the SSND problem. In particular, SSND is more difficult for smaller π. Furthermore, estimating the novelty proportion π can become arbitrarily difficult as the nominal and novel distributions become increasingly similar.

Our methodology also provides general solutions to two well-studied problems in hypothesis testing. First, our lower bound on π translates immediately to a test for $\pi = 0$, which amounts to a distribution-free solution to the two-sample problem. Second, we also show that SSND provides a powerful generalization of standard multiple testing. Important problems for future work will include developing practical methodologies for these problems based on our theoretical framework.

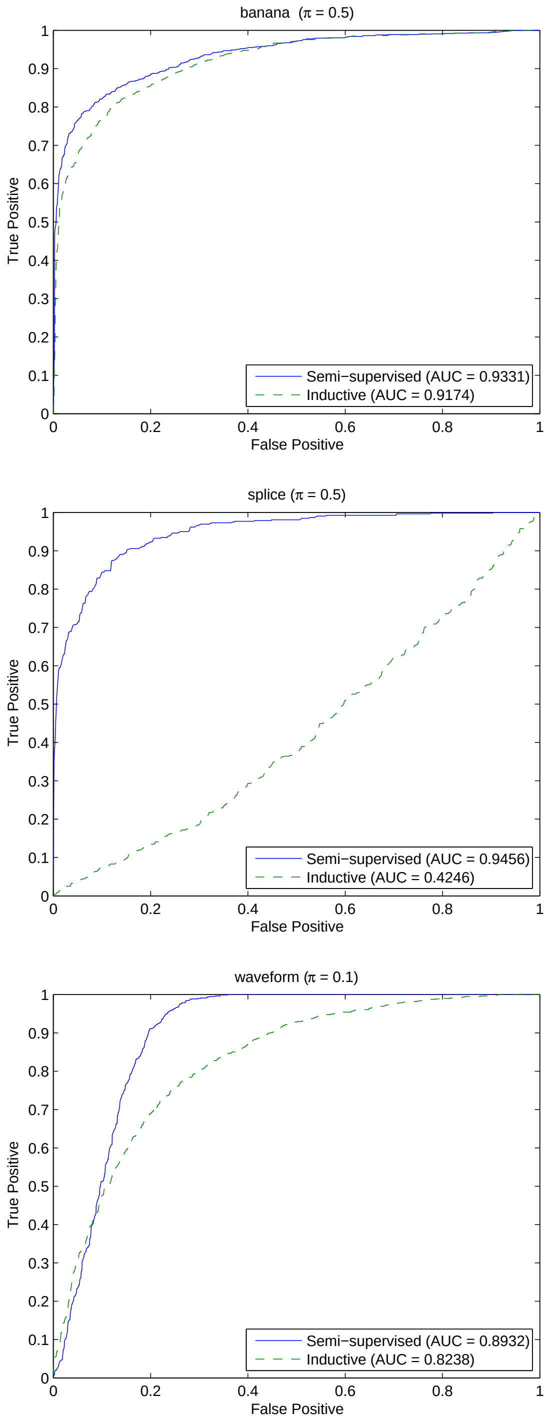

Figure 1: Illustrative results from the semi-supervised setting. Top: In the 2-dimensional banana data, the two classes are well separated, and the inductive approach fares well. Middle: In the 60-dimensional splice data, the inductive approach does worse than random guessing. Bottom: In the 21-dimensional waveform data, unlabeled data still offer gains when π is small (here 0.1).

Appendix A: Proofs

Proof of Theorem 2

For the first two claims of the theorem, we directly apply Theorem 3 of Scott and Nowak [2005] to the problem of NP classification of P_0 versus P_X , and obtain that for a suitable choice of constants c, c' we have with probability at least $1-\delta$:

$$R_0(\widehat{f_\tau}) - \alpha \leq c'\varepsilon_n \; ; \; R_X(\widehat{f_\tau}) - R_X(f^*) \leq c'\varepsilon_m \, .$$

From this, we deduce (3)-(4) by application of Theorem 1.

For the second claim, by application of Bayes' rule we have for any classifier f :

$$Q_0(f) = \frac{(1-\pi)(1-R_0(f))}{P_X(f(X)=0)} = \frac{(1-\pi)(1-R_0(f))}{\pi R_1(f) + (1-\pi)(1-R_0(f))}$$

and

$$Q_1(f) = \frac{\pi(1-R_1(f))}{P_X(f(X)=1)} = \frac{\pi(1-R_1(f))}{(1-\pi)R_0(f) + \pi(1-R_1(f))} \, .$$

Note that for $a, b > 0$ the function $(x,y) \in [0,1] \times \mathbb{R}_+ \mapsto \frac{a(1-x)}{by+a(1-x)}$ is decreasing in both variables. Hence, using (3)-(4) and the fact that $R_i(f) \in [0,1]$, we derive a lower bound on $Q_0(\widehat{f_\tau})$ as follows:

$$\begin{aligned}
Q_0(\widehat{f_\tau}) &= \frac{(1-\pi)(1-R_0(\widehat{f_\tau}))}{\pi R_1(\widehat{f_\tau}) + (1-\pi)(1-R_0(\widehat{f_\tau}))} \\
&\geq \frac{(1-\pi)(1-\alpha-c'\varepsilon_n)_+}{\pi(R_1(f^*) + c'\pi^{-1}(\varepsilon_n+\varepsilon_m)) + (1-\pi)(1-R_0(f^*)-c'\varepsilon_n)_+} \\
&\geq \frac{(1-\pi)(1-\alpha)}{P_X(f^*(X)=0) + c'(\varepsilon_m + \pi\varepsilon_n)} - \frac{c'(1-\pi)\varepsilon_n}{P_X(f^*(X)=0)} \\
&\geq \frac{(1-\pi)(1-\alpha) - c'(1-\pi)\varepsilon_n}{P_X(f^*(X)=0)} - \frac{(1-\pi)(1-\alpha)c'(\varepsilon_m+\pi\varepsilon_n)}{P_X(f^*(X)=0)^2} \\
&\geq Q_0(f^*) - \frac{c'(\varepsilon_n+\varepsilon_m)}{P_X(f^*(X)=0)} \, .
\end{aligned}$$

In the first inequality $(.)_+$ denotes the positive part. The second is elementary. In the third inequality we used the fact that the function $g : \delta \mapsto g(\delta) = \frac{A}{B+\delta}$ is convex for A, B, δ positive and has derivative $-A/B^2$ in zero, so that $g(\delta) \geq \frac{A}{B} - \delta\frac{A}{B^2}$, with $A = (1-\pi)(1-\alpha), B = P_X(f^*(X)=0), \delta = c'(\varepsilon_m + \pi\varepsilon_n)$. In the last inequality we used (with the same definition for A, B) that $\frac{A}{B} = Q_0(f^*) \leq 1$.

The treatment for Q_1 is very similar. We have

$$\begin{aligned} Q_1(\widehat{f_\tau}) &= \frac{\pi(1-R_1(f))}{(1-\pi)R_0(f)+\pi(1-R_1(f))} \\ &\geq \frac{\pi(1-R_1(f^*)-c'\pi^{-1}(\varepsilon_n+\varepsilon_m))_+}{(1-\pi)(\alpha+c'\varepsilon_n)+\pi(1-R_1(f^*)-c'\pi^{-1}(\varepsilon_n+\varepsilon_m)_+)} \\ &\geq \frac{\pi(1-R_1(f^*))}{P_X(f^*(X)=1)} - \min\left(1, \frac{c'(\varepsilon_n+\varepsilon_m)}{P_X(f^*(X)=1)-c'(\varepsilon_n+\varepsilon_m)}\right) \\ &\geq Q_1(f^*) - \frac{2c'(\varepsilon_n+\varepsilon_m)}{P_X(f^*(X)=1)}; \end{aligned}$$

we introduced the $\min(1,\ldots)$ operation since $Q_1(\widehat{f_\tau}) \geq 0$ and in the last inequality we used the fact that $\min(1, x/(1-x)) \leq 2x$ for $x \geq 0$ (with $x = c'(\varepsilon_n+\varepsilon_m)/P_X(f^*(X)=1)$).

Proof of Proposition 5

Consider the Lebesgue decomposition $P_X = P_X^0 + P_X^\perp$ with $P_X^0 \ll P_0$ (i.e. P_X^0 is absolutely continuous with respect to P_0) and $P_X^\perp \perp P_X^0$ (i.e. $P_X^\perp$ and P_0 are mutually singular). Let $f = dP_X^0/dP_0$ and a be the essential infimum of f wrt P_0. We claim that $\pi^* = 1-a$. Certainly we have the valid decomposition

$$P_X = aP_0 + (1-a)P_1, \qquad P_1 := \left((1-a)^{-1}\left((f-a)P_0 + P_X^\perp\right)\right),$$

so that $\pi^* \leq 1-a$.

By definition of singular measures there exists a measurable set D such that $P_0(D) = 1$ and $P_X^\perp(D) = 0$. Fix $\varepsilon > 0$; by definition of the essential infimum there exists a measurable set C such that $P_0(C) > 0$ and $f \leq a + \varepsilon$ P_0-a.s. on C. Put $A = C \cap D$. Then $P_0(A) = P_0(C) > 0$. Furthermore

$$\frac{P_1(A)}{P_0(A)} = \frac{E_{X\sim P_0}\left[(f-a)\mathbf{1}_{X\in A}\right]}{P_0(A)} \leq \varepsilon.$$

Existence of a decomposition of the form $P_1 = (1-\gamma)Q + \gamma P_0$ implies that for any measurable set A, $P_1(A) \geq \gamma P_0(A)$. Hence the above implies that $\gamma = 0$ for any such decomposition, i.e. P_1 must be a proper novelty distribution wrt. P_0. It also implies that for any $\varepsilon > 0$ there exists a measurable set A with $P_0(A) > 0$ and $P_X(A)/P_0(A) \leq a + \varepsilon$. By the same token, the latter implies $(1-\pi^*) \leq a$. We thus established $\pi^* = 1-a$ and the existence of the decomposition. Concerning the unicity, the decomposition established above for π^* implies that for any $\pi \geq \pi^*$, $P_X = (1-\pi)P_0 + \pi Q$ holds with $Q = (1-\frac{\pi^*}{\pi})P_0 + \frac{\pi^*}{\pi}$. Note that for any fixed π, existence of a decomposition $P_X = (1-\pi)P_0 + \pi Q$ uniquely determines Q. Hence for $\pi > \pi^*$ the corresponding Q is not a proper novelty distribution, and the only valid decomposition of P_X into P_0 and a proper novelty distribution is the one established previously.

Lemma used in Proof of Theorem 6

For the proof of Theorem 6 we made use of the following auxiliary result:

Lemma 13 *Assume P_1 is a proper novelty distribution wrt. P_0. Then for any $\varepsilon > 0$ there exists a (deterministic) classifier f such that $R_0(f) < 1$ and*

$$\frac{R_1(f)}{1 - R_0(f)} \leq \varepsilon .$$

Proof Since P_1 is a proper novelty distribution wrt. P_0, reiterating the reasoning in the proof of Proposition 5 shows that there exists a measurable set A with $P_0(A) > 0$ and $P_1(A)/P_0(A) \leq \varepsilon$. Put $\alpha = 1 - P_0(A) < 1$. Consider the classifier $f = \mathbf{1}_{A^c}$. Then $R_0(f) = P_0(f = 1) = \alpha$, while

$$0 \leq R_1(f) = P_1(f = 0) = P_1(A) \leq \varepsilon(1 - \alpha) . \tag{13}$$

This leads to the desired conclusion.

■

Proof of Theorem 8

By application of Lemma 13, for any $\varepsilon > 0$ there exists a classifier f^* such that $\frac{R_1(f^*)}{1-R_0(f^*)} \leq \varepsilon$. Then we have as in the proof of Theorem 6:

$$1 - \frac{R_X(f^*)}{1 - R_0(f^*)} = \pi \left(1 - \frac{R_1(f^*)}{1 - R_0(f^*)} \right) \geq \pi(1 - \varepsilon) .$$

Fix $\gamma > 0$ and define $\widetilde{P} = \frac{1}{2}(P_0 + P_1)$. Using the assumption of universal approximation, pick k such that there exists $f_k^* \in \mathcal{F}_k$ with $\widetilde{P}(f_k^*(X) \neq f^*(X)) \leq \gamma$. Since $\widetilde{P} \geq \frac{1}{2}P_0$ and $\widetilde{P} \geq \frac{1}{2}P_1$ this implies also $P_0(f_k^*(X) \neq f^*(X)) \leq 2\gamma$ as well as $P_X(f_k^*(X) \neq f^*(X)) \leq 2\gamma$.

From now we only work in the class $\mathcal{F}_k$ and so we omit the parameters in the notation $\varepsilon_i \equiv \varepsilon_i(\mathcal{F}_k, \delta k^{-2})$. By the union bound, the uniform control of the form (9) is valid simultaneously for all $\mathcal{F}_k$, with probability $1 - c\delta$ (with $c = \pi^2/6$). Hence with probability $1 - c\delta = 1 - c(mn)^{-2}$, we have

$$\widehat{R}_0(f_k^*) \leq R_0(f_k^*) + \varepsilon_m \leq R_0(f^*) + 2\gamma + \varepsilon_m ,$$

and also

$$\widehat{R}_X(f_k^*) \leq R_X(f_k^*) + \varepsilon_n \leq R_X(f^*) + 2\gamma + \varepsilon_n .$$

From this we deduce that with probability $1 - c(mn)^{-2}$:

$$\widehat{\pi}^-(\delta) \geq \widehat{\pi}^-(\mathcal{F}_k, (mn)^{-2}k^{-2}) \geq 1 - \frac{R_X(f^*) + 2\gamma + 2\varepsilon_n}{1 - R_0(f^*) - 2\gamma - 2\varepsilon_m} .$$

Since $\varepsilon_n, \varepsilon_m$ go to zero as $\min(m,n)$ goes to infinity we deduce that a.s. (using the Borel-Cantelli lemma, and the fact that the error probabilities are summable over $(m,n) \in \mathbb{N}^2$)

$$\liminf_{\min(m,n)\to\infty} \widehat{\pi}^-(\delta) \geq 1 - \frac{R_X(f^*)+2\gamma}{1-R_0(f^*)-2\gamma} \geq \pi(1-\varepsilon)\frac{1-R_0(f^*)}{1-R_0(f^*)-2\gamma} - \frac{4\gamma}{1-R_0(f^*)-2\gamma}.$$

Taking the limit of the above as $\gamma \to 0$ (for fixed ε and f^*), then as $\varepsilon \to 0$, leads to the conclusion.

Proof of Theorem 9

Let P_0, P_1, δ, π be given by the non-triviality assumption and P denote correspondingly the joint distribution of nominal and contaminated data. Fix some $\gamma > 0$ and a set D such that $0 < P_0(D) < \gamma$; put $A = D^c$, so that $1-\gamma < P_0(A) < 1$. Consider the distribution P_0 conditional to belonging to A, denoted $\widetilde{P}_0 = \frac{\mathbf{1}_{x\in A}}{P_0(A)}P_0$. This is a proper novelty distribution as it has it support strictly included in the support of P_0.

Consider the proper novelty distribution $\widetilde{P}_X = (1-\pi)\widetilde{P}_0 + \pi P_1$. Since it is proper, the novelty proportion of $\widetilde{P}_X$ with respect to P_0 is $\widetilde{\pi} = 1$. Finally, define the joint distribution on nominal and contaminated data $\widetilde{P} = P_0^{\otimes m} \otimes \widetilde{P}_X^{\otimes n}$.

By the non-triviality assumption, there exists a set B of (m,n) samples such that $\widehat{\pi}^+(\delta) < 1$ on the set B and $P(B) = \delta_0 > \delta$. Denote $\widetilde{A} = \mathcal{X}^m \times A^n$. By assumption, $P(\widetilde{A}) \geq (1-\gamma)^n$; furthermore by definition of $\widetilde{P}$ it can be verified straightforwardly that for any set $D \subset \widetilde{A}$, $\widetilde{P}(D) \geq P(D)$. Define now $\widetilde{B} = B \cap \widetilde{A}$; we have $P(\widetilde{B}) \geq \delta_0 - (1-(1-\gamma)^n)$. Since for all samples in $\widetilde{B}$, all points of the contaminated set belong to A, we have

$$\widetilde{P}(\widetilde{B}) \geq P(\widetilde{B}) \geq \delta_0 - (1-(1-\gamma)^n).$$

Hence for γ small enough, we have $\widetilde{P}(\widetilde{B}) > \delta$ which contradicts the fact that $\widehat{\pi}^+(\delta)$ is a $1-\delta$ confidence upper bound, since on $\widetilde{B}$ we have $\widehat{\pi}^+(\delta) < 1 = \widetilde{\pi}$.

Appendix B: Randomized classifiers and ROCs

In this appendix we recall some well-known properties of ROCs that are relevant to our setting. Let $\mathcal{F}$ be a fixed set of classifiers, and recall the Neyman-Pearson classification optimization problem (1), restated here for convenience:

$$\begin{aligned} R^*_{1,\alpha}(\mathcal{F}) := \inf_{f\in\mathcal{F}} \ & R_1(f) \\ \text{s.t.} \ & R_0(f) \leq \alpha. \end{aligned} \tag{1}$$

The optimal ROC of P_1 versus P_0 for set $\mathcal{F}$ is the function $\alpha \in [0,1] \mapsto 1 - R^*_{1,\alpha}(\mathcal{F}) \in [0,1]$.

If $\mathcal{F}$ is the set of all possible deterministic classifiers, and one assumes that both class probabilities P_0, P_1 have densities h_0, h_1 with respect to some reference measure, and such that the likelihood ratio $F(x) = \frac{h_1(x)}{h_0(x)}$ is continuous with $\inf F = 0$ and $\sup F = +\infty$, then the optimal solutions f^*_α of (1) are indicators of sets of the form

$$C_\lambda = \left\{ x \in \mathcal{X} : \frac{h_1(x)}{h_0(x)} \geq \lambda \right\},$$

with $\lambda(\alpha)$ such that $P_0(C_\lambda) = \alpha$. In this case $R^*_{1,\alpha}(\mathcal{F}) = P_1(C_{\lambda(\alpha)})$ and it can be shown that the ROC is continuous, nondecreasing and concave between the points $(0,0)$ and $(1,1)$. In particular in this case it holds that $R_0(f^*_\alpha) = \alpha$.

When some of the above assumptions are not satisfied, for example if we consider an arbitrary subset $\mathcal{F}$ of classifiers, or the probability distributions P_0 and P_1 have atoms, some of these properties may fail to hold. While it is clear that the optimal ROC is always a nondecreasing function, it might fail to be concave, and the optimal solution might have $R_0(f^*_\alpha) < \alpha$. This is for example obviously the case if $\mathcal{F}$ is a finite set of classifiers, in which case the ROC is a step function and $R_0(f)$ can only take finitely many values.

We are interested in the following regularity properties depending on $\mathcal{F}$, P_0 and P_1:

(A') For any $\alpha \in (0,1)$, there exists a sequence $f_n \in \mathcal{F}$ such that $R_0(f_n) = \alpha$ and $R_1(f_n) \to R^*_{1,\alpha}(\mathcal{F})$.

(B) The function $\alpha \mapsto R^*_{1,\alpha}(\mathcal{F})/(1-\alpha)$ is nondecreasing on $[0,1]$.

Note that for simplicity of exposition, in the main body of the paper we simplified property **(A')** into **(A)**, where the sequence f_n is replaced by its limit, assumed to belong to the considered set of classifiers. Our results still hold under **(A')** with straightforward modifications of the proofs.

Condition **(B)** states that the slope of the line joining the point of the optimal ROC at α and the point $(1,1)$ is nonincreasing in α; this is assumption is weaker than concavity of the ROC. It is relevant for the discussion in the final paragraph below, related to our result on precision.

To ensure regularity properties of the ROC, a standard device is to extend the class $\mathcal{F}$ and consider *randomized* classifiers, whose output is not a deterministic function, but a Bernoulli variable with probability depending on the point x. Formally this amounts to allowing a classifier f to take values in $[0,1]$; now for a given x the final decision $D(f,x)$ is to output class 1 with probability $f(x)$ and 0 with probability $1-f(x)$, where this decision is reached using an independent coin flipping. In this setting the error probabilities become for $y = 0,1$:

$$R_y(f) := P_y(D(f,X) \neq y) = E_y(|f(X) - y|).$$

We consider two types of extensions of a (usually deterministic) class $\mathcal{F}$, the first one is the convex hull of $\mathcal{F}$, or *full randomization,*

$$\overline{\mathcal{F}} = \left\{ g \middle| g = \sum_{i=1}^N \lambda_i f_i; N \in \mathbb{N}, f_i \in \mathcal{F}, \lambda_i \geq 0 \text{ for } 1 \leq i \leq N, \sum_{i=1}^N \lambda_i = 1 \right\}.$$

The second is given by

$$\mathcal{F}^+ = \{g | g = \lambda f + (1-\lambda), f \in \mathcal{F}, \lambda \in [0,1]\},$$

where the randomization is limited to convex interpolation between one classifier of the base class and the constant classifier equal to 1.

The following standard lemma summarizes the properties of the optimal ROC curve for these extended classes:

Lemma 14 *Let $\mathcal{F}$ be a set of deterministic classifiers containing the constant classifier equal to zero, and let P_0, P_1 be arbitrary distributions on $\mathcal{X}$. Then assumptions* **(A')** *and* **(B)** *are met when considering optimization problem* (1) *over either $\overline{\mathcal{F}}$ or $\mathcal{F}^+$. The optimal ROC for the set $\overline{\mathcal{F}}$ is concave.*

Proof The fact that the constant zero classifier belongs to $\mathcal{F}$ ensures that the infimum in (1) is not taken over an empty set and exists. Let g_n be a sequence of elements of $\mathcal{F}^+$ such that $R_0(g_n) \leq \alpha$ and $R_1(g_n) \to R^*_{1,\alpha}(\mathcal{F}^+)$. Then putting $\lambda_n = (1-\alpha)/(1-R_0(g_n))$, the sequence $f_n = \lambda_n g_n + (1-\lambda_n)$ belongs to $\mathcal{F}^+$ and ensures **(A')**. The same reasoning applies to $\overline{\mathcal{F}}$.

For property **(B)**, consider a sequence (f_n) from property **(A')**, a number $\beta \in [\alpha, 1]$ and $h_n = (1-\zeta) f_n + \zeta$ where $\zeta = (1-\beta)/(1-\alpha) \in [0,1]$. Then $h_n \in \mathcal{F}^+$, $R_0(h_n) = \beta$ and $R_1(h_n) = (1-\zeta)R_1(f_n) + \zeta$. Letting n grow to infinity we obtain $R^*_{1,\beta}(\mathcal{F}^+) \leq (1-\zeta)R^*_{1,\alpha}(\mathcal{F}^+) + \zeta$ which in turn implies $(\mathbf{B})$.

In the case of $\overline{\mathcal{F}}$, similarly consider sequences $f_{n,1}, f_{n,2}$ like above for $\alpha = \alpha_1$, resp. $\alpha = \alpha_2$ with $\alpha_2 \geq \alpha_1$; for any $\beta \in [\alpha_1, \alpha_2]$, write $\beta = \lambda\alpha_1 + (1-\lambda)\alpha_2$; correspondingly the sequence $\lambda f_{n,1} + (1-\lambda) f_{n,2}$ belongs to $\overline{\mathcal{F}}$ and ensures that $R^*_{1,\beta}(\overline{\mathcal{F}}) \leq \lambda R^*_{1,\alpha_1}(\overline{\mathcal{F}}) + (1-\lambda)R^*_{1,\alpha_2}(\overline{\mathcal{F}})$ i.e. the optimal ROC for $\overline{\mathcal{F}}$ is concave. ■

Concerning estimation error control for the extended classes, note that if a uniform error control holds over the base class $\mathcal{F}$, for example of the form (9), then the same bound holds over the extended classes $\mathcal{F}^+$ and $\overline{\mathcal{F}}$ by convex combination. Hence, for uniform statistical error control it is sufficient to consider the base class $\mathcal{F}$, for example if it is a VC-class.

For practical purposes, it might be significantly more difficult to find the solution of the (empirical version of) (1) for randomized classes and in particular for the fully randomized extension $\overline{\mathcal{F}}$. An advantage of the more limited form of randomization is that optimization problem (1) over $\mathcal{F}^+$ can be rewritten equivalently as an optimization problem over the original class, namely as

$$\inf_{h \in \mathcal{F}} \frac{R_1(h)}{1-R_0(h)} \qquad \text{s.t. } R_0(h) \leq \alpha. \tag{14}$$

To see why, assume for simplicity of exposition that **(A)** rather than **(A')** is satisfied. Then the optimization problem (1) over $\mathcal{F}^+$ is attained for some randomized classifier f^*; by construction f^* is of the form $f^* = \lambda h^* + (1-\lambda)$ for some $\lambda \in [0,1]$ and $h^* \in \mathcal{F}$. By property **(A)** we can assume $R_0(f^*) = \alpha$, which entails $\lambda = (1-\alpha)/(1-R_0(h^*))$ and $R_1(f^*) = (1-\alpha)R_1(h^*)/(1-R_0(h^*))$, hence the equivalence with (14) (with the above relation between f^* and h^*).

Finally, in general we can interpret the optimization problem (14) as a maximization of the class 0 precision,

$$Q_0(f) = P_{XY}(Y=0|f(X)=0) = \frac{(1-\pi)(1-R_0(f))}{(1-\pi)(1-R_0(f))+\pi R_1(f)} = \frac{(1-\pi)}{(1-\pi)+\pi\frac{R_1(f)}{1-R_0(f)}},$$

under the constraint $R_0(f) \leq \alpha$, since the above display shows that $Q_0(f)$ is a decreasing function of the ratio $R_1(f)/(1-R_0(f))$. In general if properties **(A)** and **(B)** are satisfied for the considered class, then it is easy to see that the solutions to (1) and (14) coincide, so that the same classifier f^* achieves the minimum FNR and class 0 precision under the constraint on the FPR.

References

C. H. Teo A. Smola, L. Song. Relative novelty detection. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS) 2009*, pages 536–543, Clearwater Beach, Florida, 2009. JMLR: W&CP 5.

S. Ben-David, T. Lu, and D. Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In R. Servedio and T. Zhang, editors, *Proc. 21st Annual Conference on Learning Theory (COLT)*, pages 33–44, Helsinki, 2008.

A. Beygelzimer, V. Dani, T. Hayes, J. Langford, and B. Zadrozny. Error-limiting reductions between classification tasks. In L. De Raedt and S. Wrobel, editors, *Proceedings of the 22nd International Machine Learning Conference (ICML)*. ACM Press, 2005.

A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the Neyman-Pearson and min-max criteria. Technical Report LA-UR 02-2951, Los Alamos National Laboratory, 2002.

Chih-Chung Chang and Chih-Jen Lin. Libsvm : a library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

Z. Chi. On the performance of FDR control: constraints and a partial solution. *Ann. Stat.*, 35(4):1409–1431, 2007.

Z. Chi. False discovery rate control with multivariate p-values. *Electronic Journal of Statistics*, 2:368–411, 2008.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Machine Learning Research*, 7:1–30, 2006.

F. Denis. PAC learning from positive statistical queries. In *Proc. 9th Int. Conf. on Algorithmic Learning Theory (ALT)*, pages 112–126, Otzenhausen, Germany, 1998.

F. Denis, R. Gilleron, and F. Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70–83, 2005.

L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans. Patt. Anal. Mach. Intell.*, 4:154–157, 1982.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

R. El-Yaniv and M. Nisenson. Optimal single-class classification strategies. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Adv. in Neural Inform. Proc. Systems 19*. MIT Press, Cambridge, MA, 2007.

C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061, 2004.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

A. Hero. Geometric entropy minimization for anomaly detection and localization. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Adv. in Neural Inform. Proc. Systems 19*. MIT Press, Cambridge, MA, 2007.

V. Kulikov and H. Lopuhaä. The behavior of the NPMLE of a decreasing density near the boundaries of the support. *Ann. Stat.*, 34(2):742–768, 2006.

J. Lafferty and L. Wasserman. Statistical analysis of semi-supervised regression. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 801–808. MIT Press, Cambridge, MA, 2008.

M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B*, 67:555–572, 2005.

W. S. Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *Proc. 20th Int. Conf. on Machine Learning (ICML)*, pages 448–455, Washington, DC, 2003.

E. Lehmann. *Testing statistical hypotheses*. Wiley, New York, 1986.

B. Liu, W. S. Lee, P. S. Yu, and X. Li. Partially supervised classification of text documents. In *Proc. 19th Int. Conf. Machine Learning (ICML)*, pages 387–394, Sydney, Australia, 2002.

B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Proc. 3rd IEEE Int. Conf. on Data Mining (ICDM)*, pages 179–188, Melbourne, FL, 2003.

K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12:181–201, Mar 2001.

P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Machine Learning Research*, 8:1369–1392, 2007.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1472, 2001.

C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Trans. Inform. Theory*, 51(8):3806–3819, 2005.

C. Scott and R. Nowak. Learning minimum volume sets. *J. Machine Learning Res.*, 7: 665–704, 2006.

C. Scott, G. Bellala, and R. Willett. The false discovery rate for statistical pattern recognition. *Electronic Journal of Statistics*, 3, 2009. 651-677.

A. Singh, R. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn't. *Proc. Neural Information Processing Systems 21* – NIPS '08, 2009.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *J. Machine Learning Research*, 6:211–232, 2005.

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

W. Sun and T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.*, 102(479):901–912, 2007.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

R. Vert and J.-P. Vert. Consistency and convergence rates of one-class SVM and related algorithms. *J. Machine Learning Research*, pages 817–854, 2006.

D. Zhang and W. S. Lee. A simple probabilistic approach to learning from positive and unlabeled examples. In *Proc. 5th Annual UK Workshop on Comp. Intell. (UKCI)*, London, UK, 2005.