Novelty detection: Unlabeled data definitely help

Clayton Scott
University of Michigan
Ann Arbor, MI, USA

Gilles Blanchard
Fraunhofer FIRST.IDA
Berlin, Germany

Abstract

In machine learning, one formulation of the novelty detection problem is to build a detector based on a training sample consisting of only nominal data. The standard (inductive) approach to this problem has been to declare novelties where the nominal density is low, which reduces the problem to density level set estimation. In this talk 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, our approach 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.

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; 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 nothing about P_X, P_0, P_1, or π, although in Section 6 we do impose a natural "resolvability" condition on the mixture P_X.

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) clas-

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

sification, 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. 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.

SSND is particularly suited to situations where the novelties lie in 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 novelties are present in the unlabeled data, the data will be more concentrated than one would expect from just the nominal component, and their presence 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 m 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.

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. Experiments are presented in Section 7, while conclusions are discussed in the final section. Proofs are presented in the text as space permits.

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 (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; estimation of the proportion of anomalies is also additionally addressed.

Classification with unlabeled: 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 (Lafferty and Wasserman, 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) do 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 $\pi = 0$ and to the estimation of π.

Multiple testing: The multiple testing problem is also concerned with the simultaneous detection of many potentially abnormal measurements (viewed as rejected null hypotheses). A frequently considered model in that framework, called the *random effects model* (see, e.g., Efron et al., 2001), is essentially identical to our contamination model. Some related ideas can be found in our proposed method for estimating the proportion of novelties and for estimating the corresponding parameter in the random effects model as in Meinshausen and Rice (2006); Donoho and Jin (2004). However, a crucial difference between this setting and SSND is that P_0 is assumed to be known in advance, and via the choice of some statistic the problem is then usually reduced to a one-dimensional setting where P_0 is uniform and P_1 is often assumed to have a concave cdf. In our setting, we don't assume any prior knowledge on the distributions, the observations are in an arbitrary space, and we attack the problem through a reduction to classification, thus introducing 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 the novelty distribution has a constant density. The assumption of uniform P_1 is effectively the approach 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. The optimal FNR for a classifier with FPR $\leq \alpha$, $0 \leq \alpha \leq 1$, is

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

where the inf is over all possible classifiers. Similarly, introduce

$$R_X(f) := P_X(f(X) = 0) = \pi R_1(f) + (1-\pi)(1-R_0(f))$$

and let

$$\begin{aligned} R^*_{X,\alpha} \quad := \quad & \inf \ R_X(f) \\ & \text{s.t. } \ R_0(f) \leq \alpha, \end{aligned} \tag{2}$$

where again the inf is over all possible classifiers. In this paper we will always assume that the infima in (1) and (2) are achieved by some classifier having exactly $R_0(f) = \alpha$ (in Section 4, we will correspondingly assume that this holds when the inf is over a class $\mathcal{F}$ of classifiers). It can be shown that this assumption is always satisfied if randomized classifiers are allowed.

By the following result, the optimal classifiers for these two problems 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. *Consider any* α, $0 \le \alpha \le 1$, *and assume* $\pi > 0$. *Let* f *be such that* $R_0(f) = \alpha$. *Then* $R_X(f) = R^*_{X,\alpha}$ *iff* $R_1(f) = R^*_{1,\alpha}$.

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

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

Proof. Suppose $R_X(f) = R^*_{X,\alpha}$ but $R_1(f) > R^*_{1,\alpha}$. Let f' be such that $R_0(f') = \alpha$ and $R_1(f') < R_1(f)$. Then

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

contradicting minimality of $R^*_{X,\alpha}$. The converse is similar, and can also be deduced from the final statement. To prove the final statement, for any f we have $R_X(f) = (1-\pi)(1-R_0(f)) + \pi R_1(f)$. Also, $R^*_{X,\alpha} = \pi R^*_{1,\alpha} + (1-\pi)(1-\alpha)$, by the first part of the theorem. By subtraction we have

$$\begin{aligned} L_{1,\alpha}(f) &= \pi^{-1}(L_{X,\alpha}(f) + (1-\pi)(R_0(f)-\alpha)) \\ &\le \pi^{-1}(L_{X,\alpha}(f) + (1-\pi)\epsilon). \quad \square \end{aligned}$$

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 $L_{X,\alpha}, \epsilon \to 0$, 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

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

where $\widehat{R}$ is the empirical version of the corresponding error quantity. Define the precision of a classifier f for class i as $Q_i(f) = P(Y = i | f(X) = i)$. 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. Let* $\mathcal{F}$ *be a set of classifier of VC-dimension* V. *Denote* f^* *the optimal classifier in* $\mathcal{F}$ *with respect to the criterion in* (1). *Fixing* $\delta > 0$ *define* $\epsilon_k = \sqrt{\frac{V\log k - \log\delta}{k}}$. *There exists absolute constants* c, c' *such that, if we choose* $\tau = c\epsilon_n$, *the following bounds hold with probability* $1-\delta$ *:*

$$\begin{aligned} R_0(\widehat{f}_\tau) - \alpha &\le c'\epsilon_n\,; \\ R_1(\widehat{f}_\tau) - R_1(f^*) &\le c'\pi^{-1}(\epsilon_n + \epsilon_m) \\ Q_i(f^*) - Q_i(\widehat{f}_\tau) &\le \frac{c'}{P(f^*(X) = i)}(\epsilon_n + \epsilon_m)\,,\ i = 0, 1\,. \end{aligned}$$

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 optimal value over the class $\mathcal{F}$. This is in contrast to the statistical learning bounds previously obtained (Liu et al., 2002, Thm. 2), which do not imply consistency. Also, following Scott and Nowak (2005), by extending suitably the argument and the method over a sequence of classes $\mathcal{F}_k$ having the universal approximation property, we can conclude that this method is universally consistent. 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.

We emphasize that the above result is but one of many possible theorems that could be deduced from the learning reduction. Other algorithms for which PAC bounds are known could easily be treated. 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: The inductive approach essentially performs density level set estimation. As shown in Steinwart et al. (2005), 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. 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 $R_0(f) \le \alpha + \epsilon$, then

$$\tilde{L}_{1,\alpha} \le \frac{1}{\tilde{\pi}}(\tilde{L}_{X,\alpha} + (1-\tilde{\pi})\epsilon) = \frac{1}{\tilde{\pi}}(\tilde{L}_{X,\alpha} + (1-\pi)(1-p_n)\epsilon). \tag{3}$$

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. *Let f be a classifier with $R_0(f) \le \alpha + \epsilon$. If $\pi = 0$, then*

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

If $\pi > 0$, then

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

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 $\epsilon = R_0(f) - \alpha$ and $\tilde{L}_{X,\alpha}$ 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 (3). To prove the second statement, denote $\beta_n := \frac{\pi(1-p_n)}{\tilde{\pi}}$, and observe that

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

Therefore

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

and we conclude

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

□

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 first treat the population case. For convenience, we assume that the support of P_1 does not entirely contain the support of P_0. This restriction can be relaxed, with some additional work, by alternately assuming that it is impossible to write $P_1 = (1-p)P_1' + pP_0$ for some P_1' and $p > 0$.

Theorem 4. *For any classifier f, we have the inequality*

$$\pi \geq \frac{1 - R_X(f) - R_0(f)}{1 - R_0(f)} \,. \tag{4}$$

Optimizing this bound over all classifiers for a fixed value of $R_0(f) = \alpha$, we obtain for any $\alpha > 0$:

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

Furthermore,

$$\pi = 1 + \left.\frac{dR^*_{X,\alpha}}{d\alpha}\right|_{\alpha=1} \,.$$

Proof. For the first inequality, just write for any classifier f

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

resulting in the inequality. For a fixed $\alpha = R_0(f(X))$, optimizing the bound over possible classifiers is equivalent to minimizing $R_X(f)$, yielding $R^*_{X,\alpha}$. By Theorem 1,

$$R^*_{X,\alpha} = (1-\pi)(1-\alpha) + \pi R^*_{1,\alpha}.$$

By assumption on the supports of P_1 and P_0, we know that $R^*_{1,\alpha} = 0$ for all $\alpha > \alpha_0$ for some α_0. Taking the derivative of both sides at 1^- establishes the result. □

The last part of this theorem suggests estimating π by estimating the slope of $R^*_{X,\alpha}$ at its right endpoint. This can be related to the problem of estimating a monotone density at its right endpoint. Rather than pursue this approach here, however, we instead employ learning-theoretic techniques to use (4) for deriving a lower confidence bound on π:

Theorem 5. *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*

$$\forall f \in \mathcal{F} \quad \left| Q(f(X)=1) - \widehat{Q}(f(X)=1) \right| \leq \epsilon_n(\mathcal{F},\delta) \,, \tag{5}$$

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

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

$$\widehat{\pi}^-(\mathcal{F},\delta) = \sup_{f\in\mathcal{F}} \frac{1 - \widehat{R}_X(f) - \widehat{R}_0(f) - (\epsilon_n + \epsilon_m)}{(1 - \widehat{R}_0(f) - \epsilon_m)_+} \,,$$

where the expression is formally defined to be $-\infty$ whenever the denominator is 0, so that the corresponding classifier is in fact discarded.

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

Proof. As in the proof of the previous lemma, 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

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

which can be solved whenever $1 - \widehat{R}_0(f) - \epsilon_m \geq 0$. □

The following result shows that $\widehat{\pi}^-(\mathcal{F},\delta)$ leads to a strongly universally consistent estimate of π. The proof relies on Theorem 5 in conjunction with the Borel-Cantelli lemma.

Theorem 6. *Consider a sequence $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of classifier sets having the 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 (5) *for* $\epsilon_n(\mathcal{F}_k, \delta) = 3\sqrt{\frac{V_k \log(n+1) - \log \delta/2}{n}}$. *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* $m, n \to \infty$.

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

Corollary 1. *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, $0 \leq \pi_0 < 1$.

7 Experiments

Despite previous work on LPUE, as discussed in Section 2, the efficacy of our proposed learning reduction has not been empirically demonstrated. To illustrate the impact of unlabeled data on novelty detection, we applied our framework to some datasets which are common benchmarks for binary classification[1]. Each dataset consists of both positive and negative examples. The negative examples were taken to be nominal. Each dataset is also divided into training and test examples. The positive training examples were not employed in the experiments. The test data were divided into two halves. The first half was used as the unlabeled data. The second half 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.

We implemented the inductive novelty detector using a thresholded kernel density estimate (KDE) with Gaussian kernel, and SSND using a plug-in KDE classifier. For each class, a single kernel bandwidth parameter was employed, and optimized by maximizing a cross-validation estimate of the area under the ROC (AUC). We emphasize that this ROC is different from the one used to evaluate the methods (see previous paragraph). In particular, it still views P_0 as the null distribution, but now the alternative distribution is taken to be P_X for SSND, and an artificial uniform distribution for the inductive detector. The label information is thus not used at any stage by SSND.

Figure 1 depicts some typical results. The top graph shows ROCs for a dataset where the two classes are fairly well separated, meaning the novelties lie in the tails of the nominal class. Thus the inductive method is close to the semi-supervised method. The middle graph represents the splice dataset, where the inductive method does worse than random guessing. To illustrate the method on different values of π, we reduced the proportion of novelties in the unlabeled data by various amounts. The bottom graph in Figure 1 shows the results for the waveform data, where the two classes also have a significant amount of overlap, when $\pi = 0.1$. We will report extensive numerical results elsewhere, including results for the hybrid approach and in the transductive setting.

8 Conclusions

We have shown that semi-supervised novelty detection reduces to Neyman-Pearson classification, thereby inheriting the properties of NP classification 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. Indeed, 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.

References

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.

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.

D. Donoho and J. Jin. Higher criticism for detecting

[1] http://ida.first.fhg.de/projects/bench/

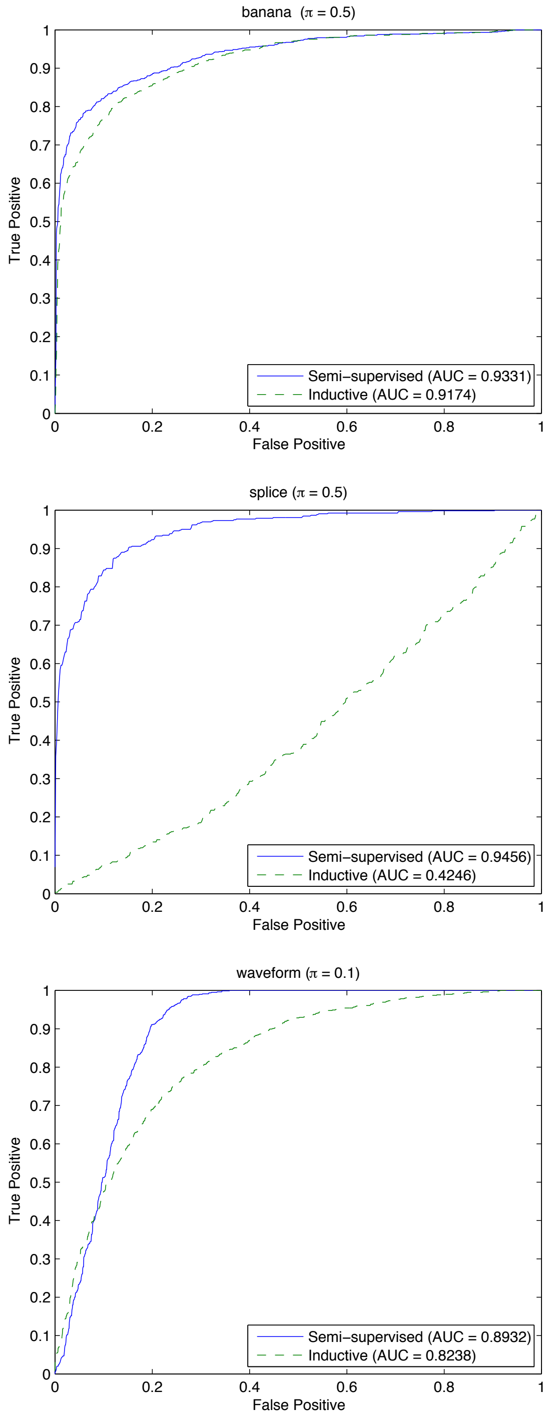

Figure 1: Top: In the banana data, the two classes are well separated, and the inductive approach fares well. Middle: In the splice data, the novelties overlap the nominal class considerably, and the inductive approach does worse than random guessing. Bottom: Unlabeled data still offer gains when π is small (here 0.1).

sparse heterogeneous mixtures. *Ann. Stat.*, 32(3):962–994, 2004.

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.

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.

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.

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.

N. Meinshausen and J. Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Stat.*, 34:373–393, 2006.

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.

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.

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.