Transductive anomaly detection

Clayton Scott and Gilles Blanchard

August 20, 2008

Abstract

One formulation of the anomaly detection problem is to build a detector based on a training sample consisting only on nominal data. The standard approach to this problem has been to declare anomalies where the nominal density is low, which reduces the problem to density level set estimation. This approach is inductive in the sense that the detector is constructed before any test data are observed. In this paper, we consider the transductive setting where the unlabeled and possibly contaminated test sample is also available at learning time. We argue that anomaly detection in this transductive setting is naturally solved by a general reduction to a binary classification problem. In particular, an anomaly detector with a desired false positive rate can be achieved through a reduction to Neyman-Pearson classification. Unlike the inductive approach, the transductive approach yields detectors that are optimal (e.g., statistically consistent) regardless of the distribution on anomalies. Therefore, in anomaly detection, unlabeled data can 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 anomaly detection. The basic task is to build a decision rule that distinguishes *nominal* from *anomalous* 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 training anomalies, however, are not available. The standard approach has been to estimate a level set of the nominal density [1, 2, 3, 4, 5], and to declare test points outside the estimated level set to be anomalies. We refer to this approach as *inductive* anomaly detection, since the decision rule is constructed before test data are observed.

In this paper we develop a *transductive* approach to anomaly detection, and argue that it 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* test sample $x_{m+1}, \ldots, x_{m+n}$ consisting potentially of both nominal and anomalous 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 anomalous ($y_i = 1$), 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 an anomaly. 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 P_1. Our specific objective is to build a decision rule with a small false negative rate subject to a fixed constraint α on the false positive rate.

Our basic contribution is to develop a general solution to the transductive anomaly detection (TAD) problem 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 TAD 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 anomaly distribution P_1, in contrast to the inductive approach which is only optimal when anomalies happen to be uniformly distributed, as discussed below. Our learning reduction allows us to import existing statistical performance guarantees for Neyman-Pearson classification [6, 7] and thereby deduce generalization error bounds, consistency, and rates of convergence for TAD.

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 (blending inductive and transductive ideas) that automatically reverts to the inductive approach when $\pi = 0$, while preserving the benefits of the NP reduction when $\pi > 0$. In addition, we discuss distribution-free one-sided confidence intervals 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 TAD. Section 5 presents our hybrid inductive/transductive approach, while Section 6 applies learning-theoretic principles to inference about π. Conclusions and future work are discussed in Section 7, and some of the longer proofs are gathered in Section 8.

2 Related work

Inductive anomaly detection: Described in the introduction, this problem is also known as one-class classification [1] or learning for only positive (or only negative) examples. The standard approach has been to assume that anomalies are outliers with respect to the nominal distribution, and to build an anomaly detector by estimating a level set of the nominal density [2, 3, 4, 5]. As we discuss below, density level set estimation implicitly assumes that anomalies are uniformly distributed. Therefore these methods can perform arbitrarily poorly (when P_1 is far from uniform), whereas the transductive approach optimally adapts to P_1.

Transductive classification: In transductive classification, labeled training data $\{(x_i, y_i)\}_{i=1}^m$ from *both* classes are given and the objective is to assign labels to the test points $\{x_i\}_{i=m+1}^{m+n}$ [8]. The setting proposed here is a special case where training data from only one class are available. Unlike the two-class problem, where unlabeled data typically do not impact theoretical properties such as consistency and rates of convergence, we argue that for anomaly detection, unlabeled data are essential for these 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 [9], 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 [10, 11] 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 considerably 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 [9, 12, 13, 14]. Most notably, Liu et al. [12] 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, [12] 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 $\pi = 0$ and to the estimation of π.

Multiple testing: The multiple testing problem is also concerned with the simultaneous detection of many anomalies (viewed as rejected null hypotheses). A frequently considered model in that framework, called the *random effects model*, (see, e.g., [15]), is essentially identical to our contamination model. Some related ideas can be found in our proposed method for estimating the proportion of anomalies and for estimating the corresponding parameter in the random effects model as in [16, 17]. However, a crucial difference between this setting and TAD is that the null distribution 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 [18], 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 [18], 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, aside from the assumption that $\pi > 0$.

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 anomaly distribution has a constant density. In other words, the inductive approach implicitly assumes anomalies are uniformly distributed. This ideas has been employed previously to reduce anomaly detection to classification by generating artifical anomalies from a uniform distribution [2, 19].

We now argue more generally, and 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

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

where the inf is over all measurable functions $f : \mathcal{X} \to \{0, 1\}$. Similarly, introduce

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

and let

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

where again the inf is over all measurable functions. 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.

The following result establishes formally the equilance between optimal tests discussed above. 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. $X \sim P_1$. This result constitutes our main *learning reduction* in the sense of [20]. 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.

Theorem 1. *Consider any α, $0 \leq \alpha \leq 1$, 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, and assume $\pi > 0$. If $R_0(f) \leq \alpha + \epsilon$, then

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

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

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

□

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. $X \sim P_X$ [6, 7], 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 [8] 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 $Q_i(f) = P(Y = i | f(X) = i)$. 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. *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). Assume $\pi > 0$ and $P(f^*(X) = i) > 0$, $i = 0, 1$. 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 $1 - \delta$:*

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

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

$$Q_i(f^*) - Q_i(\widehat{f}_\tau) \leq \frac{c'}{P_X(f^*(X) = i)}(\epsilon_n + \epsilon_m) \,, \; \textit{for } i = 0, 1 \,. \tag{5}$$

In the proof of this theorem, we show that under the constraint $R_0(f) \leq \alpha$, the best attainable precision in the set $\mathcal{F}$ for both classes is attained for $f = f^*$, so that in (5), we are really comparing the precision of $\widehat{f}_\tau$ against the best possible precision.

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 ([12], Theorem 2) for the related problem of learning from positive and unlabeled examples, which do not imply consistency. Also, following [7], 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 anomaly detection can be reduced to classification against uniform data [2], which made the statistical learning study of that problem significantly simpler.

5 The case $\pi = 0$ and a hybrid inductive/transductive approach

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 [2], 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 test points. When $\pi = 0$, the uniform points will influence the final decision rule, but when $\pi > 0$, they will be swamped by the actual anomalies.

To formalize this approach, let $0 < p_n < 1$ be a sequence of real numbers. Assume that S_0 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_0. Consider the following procedure: Let $k \sim \text{binom}(n, p_n)$. Draw k independent realizations from P_2, and redefine $x_{m+1}, \dots, x_{m+k}$ to be these values. (In practice, the uniform data would simply be appended to the test data, so that information is not erased. The present procedure, however, is slightly simpler to analyze.)

The idea now is to apply the TAD learning reduction from before to this modified test 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* anomaly. The proportion of operative anomalies in the modified test sample is $\tilde{\pi} := \pi(1 - p_n) + p_n$. The distribution of operative anomalies 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 test 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 $R_0(f) \leq \alpha + \epsilon$, then

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

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

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

If $\pi > 0$, then

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

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. The first statement follows from (6) because, when $\pi = 0$, $\tilde{L}_{1,\alpha}(f) = L_{2,\alpha}(f)$, and the right-hand side of (6) simplifies to the stated bound.

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

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

□

We remark that this hybrid procedure could be applied with any a priori distribution on anomalies besides uniform. In addition, the hybrid approach could also be practically useful when n is small, assuming the artificial points are appended to the test 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*

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

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

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

resulting in the 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. □

6.1 Distribution-free lower bounds on π

The last part of the previous 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 [21]. 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 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*

$$\forall f \in \mathcal{F} \qquad \left| Q(f(X)=1) - \widehat{Q}(f(X)=1) \right| \leq \epsilon_n(\mathcal{F}, \delta)\,, \tag{8}$$

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) :

$$\widehat{\pi}^{-}(\mathcal{F},\delta) = \sup_{f\in\mathcal{F}} \frac{1-\widehat{R}_X(f)-\widehat{R}_0(f)-(\epsilon_n+\epsilon_m)}{(1-\widehat{R}_0(f)-\epsilon_m)_+}\,. \tag{9}$$

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

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

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:

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

from which we deduce after applying the uniform bound

$$1-\widehat{R}_X(f)-\epsilon_n = \widehat{P}_X(f(X)=1)-\epsilon_n \le (1-\pi)(\widehat{R}_0(f)+\epsilon_m)+\pi\,,$$

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

Below this result is applied to testing the hypothesis $\pi=0$. 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*

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

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 (8) 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 $m,n\to\infty$.

6.2 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.

Formally, we call a *distribution-free* upper confidence bound $\widehat{\pi}^{+}(\delta)$ a function of the observed data such, for any P_0, any identifiable P_1, and any $\pi < 1$, we have $\widehat{\pi}^{+}(\delta) \geq \pi$ with probability $1-\delta$ over the draw of the two samples. By "identifiable" we mean that P_1 cannot be itself decomposed as $P_1 = \alpha P_0 + (1-\alpha)P_1'$ for some $\alpha > 0$. This ensures that π is unique. This occurs, for example, under the running assumption that the support of P_1 does not entirely contain the support of P_0.

We will show essentially that such a universal upper bound does not exist unless it is trivial. The reason is that the anomalous 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 anomalies in the contaminated data (see Corollary 2 below), but we can never be sure that there are no anomalies. This situation is similar to philosophy of hypothesis 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 $\delta > 0$ there exists a set A such that $1-\delta < P_0(A) < 1$. We call a confidence bound $\widehat{\pi}^{+}(\delta)$ *non-trivial* if there exists at least a weakly diffuse nominal distribution P_0, an anomalous distribution P_1, constants $\pi > 0$, $\delta > 0$ such that

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

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

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 1. *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 model assumptions on the generating distributions must be made. This parallels the estimation of the Bayes risk in classification [22].

6.3 Testing for $\pi = 0$

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

Corollary 2. *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 anomalies.*

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 [23], 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 (8), we have the lower bound $P_X(f(X) = 1) \geq \widehat{P}_X(f(X) = 1) - \epsilon_n$ and the upper bound $P_0(f(X) = 1) \leq \widehat{P}_0(f(X) = 1) + \epsilon_m$ (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 (9) . Furthermore, if this numerator is positive then so is the denominator since it is always larger. In the end, testing $\widehat{\pi}^{-}(\mathcal{F}, \delta) > 0$ is equivalent to testing

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

7 Conclusions

We have shown that transductive anomaly 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 TAD, as well as a lower bound and consistent estimator for π, and test for $\pi = 0$. Our transductive approach optimally adapts to the unknown anomaly distribution, unlike inductive approaches, which implicitly assume the anomalies are uniformly distributed. Indeed, our analysis strongly suggests that in anomaly detection, unlike traditional binary classification, unlabeled data are essential for attaining optimal performance in terms of tight bounds, consistency, and rates of convergence. Future work will explore learning-theoretic approaches to multiple testing, monotone density estimation, and the two-sample problem.

8 Proofs

8.1 Proof of Theorem 2

For the two first claims of the theorem, we directly apply Theorem 3 of [7] 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 $1-\delta$:

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

From this, we deduce (3)-(4) by application of Theorem 1. Note that in Theorem 1, the optimal errors R_1^* and R_X^* were defined as the best out of all possible classifiers; however it is easy to check that Theorem 1 is still valid when we restrict our attention to a fixed class $\mathcal{F}$ of classifiers and compare the errors of any $f \in \mathcal{F}$ to the the best attainable errors in that class.

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

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

and

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

Note that these relations imply, under the constraint $R_0(f) \leq \alpha$, that the best attainable precision in the set $\mathcal{F}$ for both classes is attained for $f = f^*$ (see also [24]), so we are really comparing the precision of $\widehat{f_\tau}$ against the best possible precision.

We now derive a lower bound on $Q_0(\widehat{f_\tau})$ as follows:

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

The first inequality is valid using the first two claims of the theoren, because the function $(x,y) \mapsto \frac{a(1-x)}{by+a(1-x)}$ is decreasing in both variables. The second is elementary. In the third inequality we used the fact that the function $g : \epsilon \mapsto g(\epsilon) = \frac{A}{B+\epsilon}$ is convex for A, B, ϵ positive and has derivative $-A/B^2$ in zero, so that $g(\epsilon) \geq \frac{A}{B} - \epsilon\frac{A}{B^2}$, with $A = (1-\pi)(1-\alpha), B =$

$P_X(f^*(X)=0), \epsilon = c'(\epsilon_m + \pi\epsilon_n)$. In the last inequality we used (with the same definition for A, B) that $\frac{A}{B} = Q_0(f^*) \le 1$. The treatment for Q_1 is similar.

8.2 Proof of Theorem 6

Denote S_1 the support of P_1 and $f^* = \mathbf{1}_{S_1}$. We have $P_1(f^*(X)) = 1$ and, by the assumption made on the supports at the beginning of Section 6, S_1 does not entirely contain the support S_0 of P_0 so that $0 < P_0(f^*(X)=1) =: \alpha_0$. Then we have

$$\pi = \frac{P_X(f^*(X)=1) - \alpha_0}{1-\alpha_0}.$$

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 $\widetilde{P}(f_k^*(X) \neq f^*(X)) \le \gamma$. Since $\widetilde{P} \ge \frac{1}{2}P_0$ and $\widetilde{P} \ge \frac{1}{2}P_1$ this implies also $P_0(f_k^*(X) \neq f^*(X)) \le 2\gamma$ as well as $P_X(f_k^*(X) \neq f^*(X)) \le 2\gamma$.

From now we only work in the class $\mathcal{F}_k$ and so we omit the parameters in the notation $\epsilon_i \equiv \epsilon_i(\mathcal{F}_k, \delta k^{-2})$.

By the uniform control, we have with probability $1 - c(mn)^{-2}$:

$$\widehat{P}_0(f_k^*(X)=1) \le P_0(f_k^*(X)=1) + \epsilon_m \le \alpha_0 + 2\gamma + \epsilon_m .$$

Consider now the estimated classifier $\widehat{f}$ defined as the NP-classifier at level $\alpha_0 + 2\gamma + \epsilon_m$ on class k. From the above property and the definition of $\widehat{f}$, we have with probability $1 - c(mn)^{-2}$:

$$\begin{aligned} \widehat{P}_X(\widehat{f}(X)=1) &\ge \widehat{P}_X(f_k^*(X)=1) \\ &\ge P_X(f_k^*(X)=1) - \epsilon_n \ge P_X(f^*(X)=1) - 2\gamma - \epsilon_n . \end{aligned}$$

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

$$\widehat{\pi}^-(\delta) \ge \widehat{\pi}^-(\mathcal{F}_k, (mn)^{-2}k^{-2}) \ge \frac{P_X(f^*(X)=1) - \alpha_0 - 4\gamma - 2\epsilon_m - 2\epsilon_n}{1 - \alpha_0 - 2\gamma - 2\epsilon_m - \epsilon_n}.$$

Since ϵ_n, ϵ_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 $(m,n) \in \mathbb{N}^2$)

$$\liminf_{\min(m,n)\to\infty} \widehat{\pi}^-(\delta) \ge \frac{P_X(f^*(X)=1) - \alpha_0 - 4\gamma}{1-\alpha_0-2\gamma} = \pi\frac{1-\alpha_0}{1-\alpha_0-2\gamma} - \frac{4\gamma}{1-\alpha_0-2\gamma}.$$

This is true for any $\gamma > 0$, hence the conclusion.

8.3 Proof of Theorem 7

Let P_0, P_1, δ, π be given by the non-triviality assumption. Fix some $\gamma < 0$ and a set A such 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 legitimate anomalous distribution as it has it support strictly included in the support of P_0.

Consider the fully anomalous distribution $\widetilde{P}_X = (1 - \pi)\widetilde{P}_0 + \pi P_1$. Since it is fully anomalous, the anomaly 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 $P(\widetilde{B}) \geq \delta_0 - (1-\gamma)^n$. Since for all samples in $\widetilde{B}$, all points of the contaminated set belong to A, we have

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

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 $\widehat{\pi}^+(\delta) < 1 = \widetilde{\pi}$.

References

[1] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1472, 2001.

[2] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Machine Learning Research*, vol. 6, pp. 211–232, 2005.

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

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

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

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

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

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

[9] 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.

[10] F. Denis, “PAC learning from positive statistical queries,” in *Proc. 9th Int. Conf. on Algorithmic Learning Theory (ALT)*, Otzenhausen, Germany, 1998, pp. 112–126.

[11] F. Denis, R. Gilleron, and F. Letouzey, “Learning from positive and unlabeled examples,” *Theoretical Computer Science*, vol. 348, no. 1, pp. 70–83, 2005.

[12] 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)*, Sydney, Australia, 2002, pp. 387–394.

[13] 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)*, Washington, DC, 2003, pp. 448–455.

[14] 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)*, Melbourne, FL, 2003, pp. 179–188.

[15] B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher, “Empirical Bayes analysis of a microarray experiment,” *Journal of the American Statistical Association*, vol. 96, pp. 1151–1160, 2001.

[16] N. Meinshausen and J. Rice, “Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses,” *Ann. Stat.*, vol. 34, pp. 373–393, 2006.

[17] D. Donoho and J. Jin, “Higher criticism for detecting sparse heterogeneous mixtures,” *Ann. Stat.*, vol. 32, no. 3, pp. 962–994, 2004.

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

[19] J. Theiler and D. M. Cai, “Resampling approach for anomaly detection in multispectral images,” in *Proc. SPIE*, 2003, vol. 5093, pp. 230–240.

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

[21] V. Kulikov and H. Lopuhaä, "The behavior of the NPMLE of a decreasing density near the boundaries of the support," *Ann. Stat.*, vol. 34, no. 2, pp. 742–768, 2006.

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

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

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