Journal of Machine Learning Research 23 (2022) 1-47 Submitted 4/21; Revised 9/21; Published 2/22

PAC Guarantees and Effective Algorithms for Detecting Novel Categories

Si Liu LIUS2@OREGONSTATE.EDU
Department of Statistics
Oregon State University
Corvallis, OR 97331-4606, USA

Risheek Garrepalli GARREPAR@OREGONSTATE.EDU
School of EECS
Oregon State University
Corvallis, OR 97331-5501, USA

Dan Hendrycks HENDRYCKS@BERKELEY.EDU
School of EECS
University of California
Berkeley, CA 94720-1776, USA

Alan Fern ALAN.FERN@OREGONSTATE.EDU
School of EECS
Oregon State University
Corvallis, OR 97331-5501, USA

Debashis Mondal MONDAL@WUSTL.EDU
Department of Mathematics and Statistics
Washington University
St. Louis, MO 63130-4899, USA

Thomas G. Dietterich TGD@CS.ORST.EDU
School of EECS
Oregon State University
Corvallis, OR 97331-5501, USA

Editor: Mehryar Mohri

Abstract

Open category detection is the problem of detecting "alien" test instances that belong to categories or classes that were not present in the training data (Liu et al., 2018). In many applications, reliably detecting such aliens is central to ensuring the safety and accuracy of test set predictions. Unfortunately, there are no algorithms that provide theoretical guarantees on their ability to detect aliens under general assumptions. Further, while there are algorithms for open category detection, there are few empirical results that directly report alien detection rates. Thus, there are significant theoretical and empirical gaps in our understanding of open category detection. In this paper, we take a step toward addressing this gap by studying a simple, but practically-relevant variant of open category detection. In our setting, we are provided with a "clean" training set that contains only the target categories of interest and an unlabeled "contaminated" training set that contains a fraction α of alien examples. Under the assumption that we know an upper bound on α,

we develop an algorithm that gives PAC-style guarantees on the alien detection rate, while aiming to minimize false alarms. Given an overall budget on the amount of training data, we also derive the optimal allocation of samples between the mixture and the clean data sets. Experiments on synthetic and standard benchmark datasets evaluate the regimes in which the algorithm can be effective and provide a baseline for further advancements. In addition, for the situation when an upper bound for α is not available, we employ nine different anomaly proportion estimators, and run experiments on both synthetic and standard benchmark data sets to compare their performance.

Keywords: open category detection, anomaly detection, alien detection rate, false positive rate, PAC guarantees

1. Introduction

Most machine learning systems implicitly or explicitly assume that their training experience is representative of their test experience. This assumption is rarely true in real-world deployments of machine learning, where "unknown unknowns", or "alien" test queries, can arise without warning. Ignoring the potential for such aliens can lead to serious safety concerns in many applications and significantly degrade the accuracy of test set predictions in others. For example, consider a scientific application (Lytle et al., 2010) where a classifier is trained to recognize specific categories of insects in freshwater samples in order to detect important environmental changes. Test samples will typically contain some fraction of specimens belonging to species not represented in the training data. A classifier that is unaware of these new species will misclassify the specimens as belonging to existing species. This will produce incorrect conclusions.

The problem of open category detection is to detect such alien examples at test time. The primary approach to open category detection is to train an anomaly detector based on a "clean" training set (one that contains no aliens). The anomaly detector, when applied to a query instance x, returns an anomaly score $A(x)$. This score is compared against a threshold τ, and if $A(x) > \tau$, then an alarm is raised that declares that x is an alien. The open category detection performance will depend on both the quality of the anomaly detector and the setting of the alarm threshold τ. Ideally, the anomaly scores of all alien instances are larger than the scores of all nominal (known category) instances, and τ can be set to perfectly separate the aliens from the nominals. Unfortunately, this is rarely achieved in practice. Figure 1 illustrates a more typical case in which the anomaly score distributions overlap. In such cases, we confront a tradeoff between false alarms (nominal cases falsely declared to be aliens) and missed alarms (aliens incorrectly declared to be nominal), and we must set τ to achieve an operating point along this tradeoff. If we had additional labeled data for nominals and aliens, we could estimate the anomaly score distributions shown in Figure 1 and choose τ. In most applications, however, we do not have labeled aliens. Indeed, this is why the task of open category detection arises in the first place.

Prior research has focused on controlling only the false alarm rate. Given labeled nominal data, we can estimate the distribution of anomaly scores for the nominal instances and select τ to achieve a desired false alarm rate, as shown by τ_{FAR} in Figure 1. The central question in this paper is whether we can instead set τ to control the missed alarm rate, as indicated by τ_{MAR} in Figure 1. Although both false alarms and missed alarms are impor-

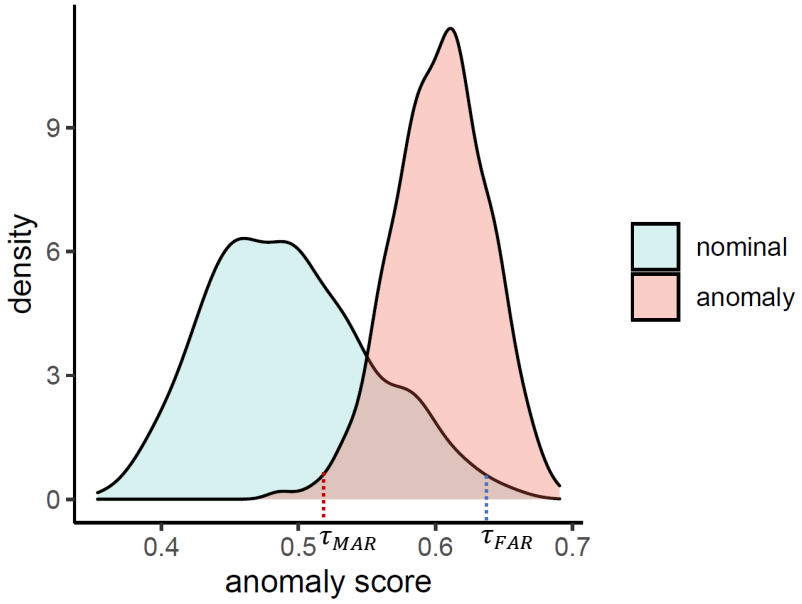

Figure 1: Density curves of anomaly scores for nominal and anomalous data points. Setting $\tau := \tau_{FAR}$ controls the area under the upper tail of the nominal density. Setting $\tau := \tau_{MAR}$ controls the area under the lower tail of the anomaly density.

tant, controlling missed alarms is much more relevant to safety-critical applications. In this paper, we present a method that can provide PAC guarantees on the missed alarm rate.

In the absence of labeled anomaly data, our approach assumes two training sets: a labeled clean training data set with labels drawn from a finite set of known categories and an unlabeled contaminated data set that contains a fraction α of queries that belong to unknown categories (alien queries). Our first contribution is to show that, in this setting, theoretical guarantees are possible given knowledge of an upper bound on α. In particular, we give an algorithm that employs this knowledge to provide Probably Approximately Correct (PAC) guarantees for achieving a user-specified alien detection rate. While knowledge of a non-trivial upper bound on α may not always be possible, in many situations it will be possible to select a reasonable value based on domain knowledge, prior data, or by inspecting a sample of the test data.

The utility of our method depends on the quality of the anomaly detector. The detection rate guarantee will be met regardless of the quality of the anomaly detector. But a poor anomaly detector will exhibit a very high false alarm rate when τ is set by our algorithm.

Because our method relies on having two different kinds of training data, the question arises of how the performance depends on the relative sizes of those data sets. To answer this question, we analyze the case where we are given a fixed budget on the total size of the two data sets. We determine the optimal way of dividing this budget between the two samples. We show that as α grows large, the size of the unlabeled mixture sample should be larger than the labeled data set.

We carry out experiments[1] on synthetic and benchmark data sets using a state-of-the-art anomaly detector, the Isolation Forest invented by Liu et al. (2008). We vary the amount of training data, the fraction α of alien data points, along with the tightness of the upper bound on α. The results indicate that our algorithm can achieve the guaranteed performance when enough data is available, as predicted by the theory. The results also show that for the considered benchmarks, the Isolation Forest anomaly detector is able to support non-trivial false positive rates given enough data. The results also illustrate the inherent difficulty of the problem for small data sets and/or small values of α.

In practice, there can be many situations where an upper bound on α is not available. There have been multiple studies on estimating α in literature. We evaluate those methods on both synthetic and UCI benchmark data sets and compare them to a new estimation method that we introduce. The experiments show that several of the α estimation methods give reasonable estimates, and the performance of open category detection using those estimates gives results that will be usable in practice.

This work is an extension of a previous conference version by Liu et al. (2018).

2. Related Work

The problem of open category detection is related to several other lines of research in machine learning and statistics.

Open category detection is related to the problem of one-class classification, which aims to detect outliers relative to a single training class. One-class SVMs (OCSVMs) proposed by Schölkopf et al. (2001) are popular for this problem. However, they have been found to perform poorly for open category detection due to poor generalization as in the work of Zhou and Huang (2003), which has been partly addressed by later work of Manevitz and Yousef (2002), Wu and Ye (2009), Jin et al. (2004) and Cevikalp and Triggs (2012). OCSVMs have been employed by Heflin et al. (2012) and Pritsos and Stamatatos (2013) in a multi-class setting similar to open category detection. However, there are no direct mechanisms to control the alien detection rate of these methods, which is the primary goal in our problem setting.

Menon and Williamson (2018) show that calibrated anomaly detection could be achieved through minimization of a suitable proper loss from binary classification. But the calibration is in terms of the density level sets from the mixture distribution, which does not give direct control over the alien detection rate.

Work on classification with rejection/abstaining options (Chow, 1970; Wegkamp, 2007; Tax and Duin, 2008; Pietraszek, 2005; Cortes et al., 2016; Geifman and El-Yaniv, 2017) allows classifiers to abstain from making predictions when they are not confident. While loosely related to open category detection, these approaches do not directly consider the possibility of novel categories, but rather focus on assessing confidence with respect to the known categories. Due to their closed-world discriminative nature, it is easy to construct scenarios where such methods are incorrectly confident about the class of an alien and do not abstain.

A variety of prior work has addressed variants of open category detection. This includes work by Scheirer et al. (2013) formalizing the concept of "open space" to characterize the

1. Code for reproducing our experiments can be found at `https://github.com/liusi2019/ocd-journal`.

region of the feature space outside of the support of the training set. Variants of SVMs have also been developed, such as the One-vs-Set Machine by Scheirer et al. (2013) and the Weibull-calibrated SVM by Scheirer et al. (2014). Da et al. (2014) have addressed open category detection by tuning the decision boundary based on unlabeled data that contains data from novel categories. Júnior et al. (2017) propose an approach based on nearest neighbor methods. None of these methods, however, allow for the direct control of alien detection rates, nor do they provide theoretical guarantees.

This problem also goes by several other names in the literature. Blanchard et al. (2010) frame it as a "semi-supervised novelty detection" problem, view the problem from a Neyman-Pearson classification perspective, show that the optimal test for clean against alien is identical to testing for clean against mixture, and provide guarantees on the performance of a constrained empirical risk minimizer. But in their work the goal is to maximize the recall (alien detection rate), subject to a constraint on the false alarm rate. While in our work we aim at getting the smallest possible false alarm rate, under the constraint that we should achieve a target missed alarm rate. These are two different directions, and the nature of the problems are different. Sanderson and Scott (2014) treat the open category problem as one version of the "domain adaptation" problem and show that under certain conditions, a sequence of approximate empirical risk minimizers have their risks converging in probability to the corresponding Bayes error. This is a consistency result and does not give an explicit finite sample guarantee.

There is also recent interest in open category detection for deep neural networks applied to vision and text classification (e.g., Bendale and Boult, 2016; Shu et al., 2017). These methods usually train a neural network in a standard closed-world setting, but then analyze various activations in the network in order to detect aliens. Dhamija et al. (2018) include "known unknown" classes during the training and improve the network robustness towards out of distribution samples by increasing entropy and decreasing magnitude for the unknown inputs. Hendrycks et al. (2018) fine-tune pre-trained classifiers using out-of-distribution samples and show that the resulting anomaly detectors generalize well for detecting unseen anomalies. Another related line of work is detection of out-of-distribution instances, which is similar to open category detection but assumes that the test data come from a completely different distribution compared to the training distribution (e.g., Hendrycks and Gimpel, 2017; Liang et al., 2018). All of this work is quite specialized to deep neural networks and does not provide direct control of alien detection rates or theoretical guarantees.

3. Problem Setting

We consider open category detection where there is an unknown nominal data distribution D_0 over labeled examples from a known set of category labels. We receive as input a "clean" nominal training set S_0 containing k i.i.d. draws from D_0. In practice, S_0 will correspond to some curated labeled data that contains only the known categories of interest.

We also receive as input an unlabeled "mixture" data set S_m that contains n points drawn i.i.d. from a mixture distribution D_m. Specifically, the mixture distribution D_m is a combination of the nominal distribution D_0 and an unknown alien distribution D_a, which is a distribution over novel categories (alien data points). We assume that D_a is stationary,

so that all alien points that appear as future test queries will also be drawn from D_a. (We discuss this assumption in Section 9.)

At training time, we assume that D_m is a mixture distribution, with probability α of generating an alien data point from D_a and probability of $1 - \alpha$ of generating a nominal point. Our results hold even if the test queries come from a mixture with a different value of α as long as the alien test points are drawn from D_a.

Given these data sets, our problem is to label test instances from D_m as either "alien" or "nominal". In particular, we wish to guarantee a specified *alien detection rate* (recall, or 1 minus the missed alarm rate), which is the fraction of alien data points in D_m that are classified as "alien" (e.g., 95%). At the same time we would like the *false positive rate* (false alarm rate) to be small, which is the fraction of nominal data points incorrectly classified as aliens.

Our approach to this problem assumes the availability of an anomaly detector that is trained on S_0 and assigns anomaly scores to all data points in both S_0 and S_m. The anomaly scores determine an order over the test examples according to how anomalous they appear relative to the nominal data (higher scores being more anomalous). An ideal detector would rank all alien data points higher than all nominals, though in practice, the ordering will not be so clean. Our approach labels data in S_m by selecting a threshold τ on the anomaly scores and labeling all data points with scores above the threshold as aliens and the remaining points as nominals. Our key challenge is to select a value for τ that provides a guarantee on the alien detection rate. Note that we do not provide a guarantee on the false positive rate, and the actual false positive rate that is achieved will depend on how well the anomaly detection method separates the true positives (aliens) and true negatives (nominals).

4. Algorithms for Open Category Detection

In order to obtain a theoretical guarantee, our algorithm assumes knowledge of the alien mixture probability α that generates the mixture data S_m. Later, we will show that knowing an upper bound on α is sufficient to obtain a guarantee, and we will introduce practical methods for estimating α.

Our approach is based on considering the cumulative distribution functions (CDFs) over anomaly scores of a fixed anomaly detector. Let F_0, F_a, and F_m be the CDFs of anomaly scores for the nominal data distribution D_0, alien distribution D_a, and mixture distribution D_m respectively. Since D_m is a simple mixture of D_0 and D_a, we can write F_m as

$$F_m(x) = (1 - \alpha)F_0(x) + \alpha F_a(x).$$

From this we can derive the CDF for F_a in terms of F_m and F_0:

$$F_a(x) = \frac{F_m(x) - (1 - \alpha)F_0(x)}{\alpha}.$$

Given the ability to derive F_a, it is straightforward to achieve an alien detection rate of $1 - q$ (e.g. 95%) by selecting an anomaly score threshold τ_q that is the q quantile of F_a and raising an alarm on all test queries whose anomaly score is greater than τ_q.

In reality, we do not have access to F_m or F_0 and hence cannot exactly determine F_a. Rather, we have samples S_m and S_0. Thus, our algorithm works with the empirical CDFs $\hat{F}_0$ and $\hat{F}_m$, which are simple step-wise constant approximations, and estimates an empirical CDF over aliens:

$$\hat{F}_a(x) = \frac{\hat{F}_m(x) - (1-\alpha)\hat{F}_0(x)}{\alpha}. \tag{1}$$

Our algorithm computes the above estimate of $\hat{F}_a$ and uses it to select a threshold $\hat{\tau}_q$ to be the largest threshold such that $\hat{F}_a(\hat{\tau}_q) \leq q$, where $1-q$ is the target alien detection rate. This choice will minimize the number of false alarms.

The steps of this algorithm are as follows.

Algorithm 1 Estimate Threshold

1: Compute anomaly scores for all points in S_0 and S_m, denoted $x_1, x_2, \ldots, x_k$ and $y_1, y_2, \ldots, y_n$ respectively.
2: Compute empirical CDFs $\hat{F}_0$ and $\hat{F}_m$.
3: Calculate $\hat{F}_a$ using equation 1.
4: Output detection threshold

$$\hat{\tau}_q = \max\{u \in S : \hat{F}_a(u) \leq q\},$$

where $S = \{x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_n\}$.

Although $\hat{F}_m$ and $\hat{F}_0$ are both legal CDFs, the estimate for $\hat{F}_a$ from step 3 may not be a legal CDF, because it is the difference of two noisy estimates—it may not increase monotonically, and it may even be negative. A good technique for dealing with this problem is to employ isotonization (Barlow and Brunk, 1972) and clipping. Isotonization finds the monotonically increasing function $\hat{F}_a^*$ closest to $\hat{F}_a$ in squared error. To convert $\hat{F}_a$ into a legal CDF, define $\check{F}_a = \min\{\max\{\hat{F}_a^*, \mathbf{0}\}, \mathbf{1}\}$, where the min and max operators are applied pointwise to their arguments. We performed experiments (shown in the Appendix) to test whether using $\check{F}_a$ in Step 4 would improve the performance of the overall algorithm. We found that it did not.

5. Finite Sample Guarantee

In the limit of infinite data (both nominal and mixture) and perfect knowledge of α, $\hat{F}_a$ will converge to the true alien CDF, and our algorithm will achieve the desired alien detection rate. In this section, we consider the finite data case where $|S_0| = |S_m| = n$. We derive a value for the sample size n that guarantees with high probability over random draws of S_0 and S_m that fraction $1 - q - \epsilon$ of the alien test points will be detected, where ϵ is an additional error incurred because of the finite sample size n.

Our key theoretical tool is a finite sample result due to Massart (1990) on the uniform convergence of empirical CDF functions. To use this result, we make the reasonable technical assumption that the nominal and alien CDFs, F_0 and F_a, are continuous. In the following, let η be the target alien detection rate, q be the input to Algorithm 1, $\hat{\tau}_q$ be the estimated q-quantile of the alien CDF (step 4 of Alg. 1), and ϵ be an error parameter.

The following theorem gives the sample complexity for guaranteeing that $1-\eta$ of the alien examples will be detected using threshold $\hat{\tau}_q$.

Theorem 1 *Let S_0 and S_m be nominal and mixture data sets containing n i.i.d. samples from the nominal and mixture data distributions respectively. For any $\epsilon \in (0, 1-q)$ and $\delta \in (0,1)$, if*

$$n > \frac{1}{2} \ln \frac{2}{1-\sqrt{1-\delta}} \left(\frac{1}{\epsilon}\right)^2 \left(\frac{2-\alpha}{\alpha}\right)^2,$$

then with probability at least $1-\delta$, Algorithm 1 will return a threshold $\hat{\tau}_q$ that achieves an alien detection rate of at least $1-\eta$, where $\eta = q + \epsilon$.

The proof is in the Appendix. Note that n grows as $O(\frac{1}{\epsilon^2\alpha^2} \log \frac{1}{\delta})$. Hence, this guarantee is polynomial in all relevant parameters, which we believe is the first such guarantee for open category detection. The result can be generalized to the case where $n_0 < n_m$; in practice, the larger the mixture sample S_m is, the easier it is to estimate τ_q, because this provides more alien points for estimating the q-th quantile of F_a.

The theorem suggests what value we should choose for q. For a given α, under a fixed sample size $|S_0| = |S_m| = n$ and a probability tolerance δ, the smallest ϵ we can achieve is $\frac{2-\alpha}{\alpha}\sqrt{\frac{1}{2n} \ln \frac{2}{1-\sqrt{1-\delta}}}$. If a recall of at least $1-\eta$ is desired, we should choose $q = \eta - \epsilon$. By choosing a q value smaller than $\eta - \epsilon$, we can achieve a recall greater than $1-\eta$, but this will have a higher false positive rate.

The ϵ captures how close our CDF estimate $\hat{F}_a$ is to the true CDF F_a and how much worse than $1-q$ our recall will be. As long as $\epsilon < \eta$, we can achieve the desired recall $1-\eta$ by setting $q = \eta - \epsilon$. But if we can achieve a smaller value for ϵ, we can use a larger value for q, which allows us to potentially choose a larger threshold $\hat{\tau}_q$. And the greater $\hat{\tau}_q$ is, the smaller the false positive rate will be. In the extreme case when $\epsilon \to 0$ and $n \to \infty$, $\hat{\tau}_q \to \tau_q$ and the false positive rate will tend to $1 - F_0(\tau_q)$.

The achievable false positive rate under a specified recall is decided by the relative shapes and locations of the distributions given by F_a and F_0, which are determined by the performance of the anomaly detector and the difficulty of the anomaly detection problem (i.e., how different the aliens and nominals are). Figure 2 shows one example of anomaly scores from an Isolation Forest computed on the UCI Shuttle data set. In this figure, the distributions of anomaly scores from nominal and anomalous data points are well separated, and we can attain a small false positive rate while achieving high recall. Figure 3 shows the corresponding anomaly score distributions on the Optical Recognition of Handwritten Digits data set. Here the distributions are not as well separated as those in Figure 2. If we want to achieve high recall, we are going to have a relatively high false positive rate as well. The greater $\hat{\tau}_q$ is, the smaller the resulted false positive rate will be.

Making use of the result of Massart (1990), once we have an threshold estimate $\hat{\tau}_q$, we can get a guarantee for its performance in terms of the false alarm rate as well. Under the same technical assumption of Theorem 1, the following corollary gives a sample complexity when we want both a guarantee on recall and an estimate for the resulting false positive rate.

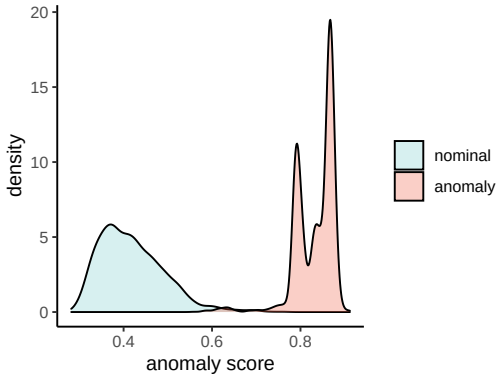

Figure 2: Density curves of anomaly scores of nominal and anomalous points from the shuttle data set in UCI repository.

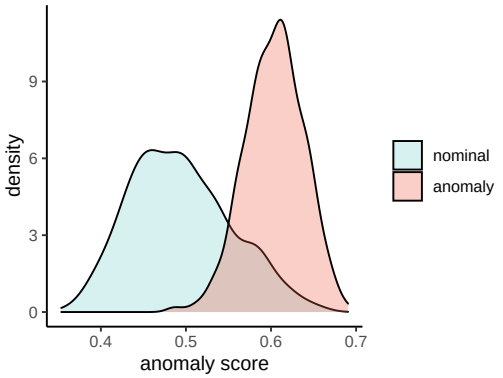

Figure 3: Density curves of anomaly scores of nominal and anomalous points from the OCR data set in UCI repository.

Corollary 1 *Under the same setting of Theorem 1, for any* $\epsilon \in (0, 1-q)$ *and* $\delta \in (0,1)$*, if*

$$n > \frac{1}{2} \ln \frac{2}{1-\sqrt{1-\delta/2}} \left(\frac{1}{\epsilon}\right)^2 \left(\frac{2-\alpha}{\alpha}\right)^2,$$

then with probability at least $1-\delta$*, Algorithm 1 will return a threshold* $\hat{\tau}_q$ *that achieves an alien detection rate of at least* $1-\eta$*, where* $\eta = q+\epsilon$*, and the false positive rate from using* $\hat{\tau}_q$ *will be no greater than* $1-\hat{F}_0(\hat{\tau}_q)+\epsilon_0$*, where* $\epsilon_0 = \sqrt{\frac{1}{2n}\ln\frac{4}{\delta}} < \epsilon$ *and* $\hat{F}_0$ *is the empirical CDF of the anomaly scores of the nominal data set only.*

The proof is in the appendix. Note that the sample size required here is larger than that in Theorem 1, due to the need to splitting the probability tolerance δ between the guarantee on recall and the estimation for the resulting false positive rate. Also note that here we are estimating the false positive rate introduced by our threshold instead of trying to find a threshold that can achieve a desired false positive rate. The requirement to control the recall and false positive rate at the same time is not always feasible, due to the relative shape of F_0 and F_a.

What if we don't know the exact value of α? If our algorithm uses an upper bound α' on the true α to compute $\hat{F}_a$, we can still provide a guarantee. In this case, in addition to the assumptions in Theorem 1, we need a concept of an anomaly detector being *admissible*. We say that an anomaly detector is *admissible* for a problem if the anomaly score CDFs satisfy $F_0(x) \geq F_m(x)$ for all $x \in \mathbb{R}$. Most reasonable anomaly detectors will be admissible in this sense, since the alien CDF will typically concentrate more mass toward larger anomaly score values compared to F_0. Indeed, if this is not the case, there is little hope, since there is effectively no signal to distinguish between aliens and nominals. Both Figure 2 and Figure 3 show examples of anomaly scores from admissible anomaly detectors.

Corollary 2 *Consider running Algorithm 1 using an upper bound α' on the true α. Under the same assumptions as Theorem 1, if the anomaly detector is admissible and*

$$n > \frac{1}{2} \ln \frac{2}{1 - \sqrt{1-\delta}} \left(\frac{1}{\epsilon}\right)^2 \left(\frac{2-\alpha'}{\alpha'}\right)^2,$$

then with probability at least $1-\delta$, Algorithm 1 will return a threshold $\hat{\tau}_q$ that achieves an alien detection rate of at least $1-\eta$, where $\eta = q + \epsilon$.

The proof is in the Appendix. While we can achieve a guarantee using an upper bound on α', the returned threshold will be more conservative (smaller) than if we had used the true α. This will result in higher false alarm rates, since more nominal points will be above the threshold. Thus it is desirable to use a value of α' that is as close to α as possible.

6. Optimal Allocation of Total Sample Size Budget

In the preceding section, we discussed the sample size requirement under the constraint $n_m = n_0$, where $n_m = |S_m|$ and $n_0 = |S_0|$. What if we relax this constraint? What will be the best way of allocating samples if we only have a constraint on the total budget $n = n_0 + n_m$?

The closer $\hat{F}_a$ is to F_a, the better the threshold estimate we can get from the algorithm. The best way of allocating samples should minimize the maximum absolute distance ϵ between $\hat{F}_a(x)$ and $F_a(x)$, subject to the requirement on δ and n, under a given α. Since δ appears inside the logarithm, its effect on the sample size requirement is limited. To facilitate the analysis, we relax the requirement on the independence between the clean sample and mixture sample. From the proof of Theorem 1, we can see that this corresponds to adopting the constraint $\delta_0 + \delta_m \leq \delta$ instead of $(1-\delta_0)(1-\delta_m) \leq 1-\delta$, where δ_0 and δ_m are the probability error tolerances we allow for estimating F_0 and F_m respectively. We can also see from the same proof that given n_0, n_m, δ_0, and δ_m, the smallest ϵ value we can achieve under a $1-\delta$ probability guarantee is

$$\epsilon(n_0, n_m, \delta_0, \delta_m) = \frac{1}{\alpha} \frac{1}{\sqrt{n_m}} \sqrt{\frac{1}{2} \ln \frac{2}{\delta_m}} + \frac{1-\alpha}{\alpha} \frac{1}{\sqrt{n_0}} \sqrt{\frac{1}{2} \ln \frac{2}{\delta_0}}.$$

The optimization problem we want to solve is therefore:

$$\begin{array}{ll} \text{minimize} & \epsilon(n_0, n_m, \delta_0, \delta_m) \\ \text{subject to} & n_0 + n_m \leq n, \\ & \delta_0 + \delta_m \leq \delta, \\ & n_0, n_m, \delta_0, \delta_m > 0. \end{array} \tag{2}$$

Given a fixed pair (δ_0, δ_m), for any pair of (n_0, n_m) satisfying $n_0 + n_m < n$, if we increase either n_0 or n_m, the value of ϵ will decrease. A similar observation holds for any pair of (δ_0, δ_m) given a fixed pair (n_0, n_m). Hence the minimum value of $\epsilon(n_m, n_0, \delta_m, \delta_0)$ can only be achieved when first two inequalities in (2) become equalities. Hence, we only need to consider the feasible region where $n_0 = n - n_m$ and $\delta_0 = \delta - \delta_m$. Thus we can simplify the optimization problem (2) to be

$$\begin{array}{ll} \text{minimize} & \epsilon(n_m, \delta_m) \\ \text{subject to} & 0 < n_m < n, \\ & 0 < \delta_m < \delta, \end{array}$$

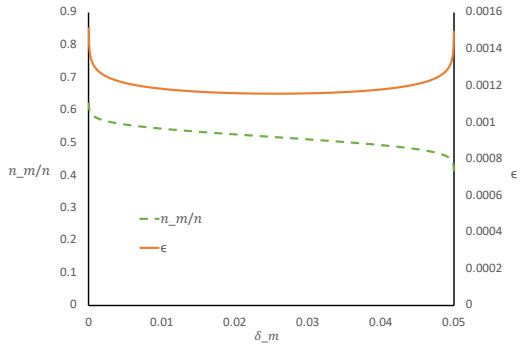

Figure 4: The optimal proportion n_m/n and corresponding optimal ϵ v.s δ_m, given $\delta = 0.05$, $\alpha = 0.10$ and $n = 1\,000\,000\,000$.

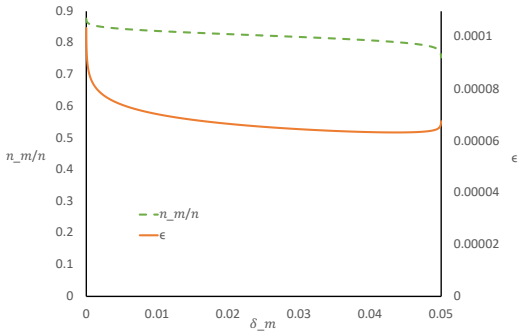

Figure 5: The optimal proportion n_m/n and corresponding optimal ϵ v.s δ_m, given $\delta = 0.05$, $\alpha = 0.90$ and $n = 1\,000\,000\,000$.

where

$$\epsilon(n_m, \delta_m) = \frac{1}{\alpha}\frac{1}{\sqrt{n_m}}\sqrt{\frac{1}{2}\ln\frac{2}{\delta_m}} + \frac{1-\alpha}{\alpha}\frac{1}{\sqrt{n-n_m}}\sqrt{\frac{1}{2}\ln\frac{2}{\delta-\delta_m}}.$$

Theorem 2 *The objective function $\epsilon(n_m, \delta_m)$ is convex in n_m and δ_m, $n_m \in (0, n)$, $\delta_m \in (0, \delta)$. Further, if we write*

$$\epsilon(\delta_m) := \min_{n_m \in (0,n)} \epsilon(n_m, \delta_m),$$

then $\epsilon(\delta_m)$ is also a convex function in δ_m.

For a fixed $\delta_m \in (0, \delta)$, $\epsilon(n_m, \delta_m)$ is a convex function in n_m, and we can analytically minimize $\epsilon(n_m, \delta_m)$ with respect to n_m to obtain

$$n_m = \frac{(1-\alpha)^{-\frac{2}{3}}\left(\frac{\ln\frac{2}{\delta_m}}{\ln\frac{2}{\delta-\delta_m}}\right)^{\frac{1}{3}}}{1+(1-\alpha)^{-\frac{2}{3}}\left(\frac{\ln\frac{2}{\delta_m}}{\ln\frac{2}{\delta-\delta_m}}\right)^{\frac{1}{3}}} n.$$

The optimal ϵ that can be achieved by each value of $\delta_m \in (0, \delta)$ is

$$\epsilon(\delta_m) = \frac{1}{\alpha}\frac{1}{\sqrt{\frac{(1-\alpha)^{-\frac{2}{3}}\left(\frac{\ln\frac{2}{\delta_m}}{\ln\frac{2}{\delta-\delta_m}}\right)^{\frac{1}{3}}}{1+(1-\alpha)^{-\frac{2}{3}}\left(\frac{\ln\frac{2}{\delta_m}}{\ln\frac{2}{\delta-\delta_m}}\right)^{\frac{1}{3}}} n}}\sqrt{\frac{1}{2}\ln\frac{2}{\delta_m}} + \frac{1-\alpha}{\alpha}\frac{1}{\sqrt{\frac{1}{1+(1-\alpha)^{-\frac{2}{3}}\left(\frac{\ln\frac{2}{\delta_m}}{\ln\frac{2}{\delta-\delta_m}}\right)^{\frac{1}{3}}} n}}\sqrt{\frac{1}{2}\ln\frac{2}{\delta-\delta_m}}.$$

Note that the optimal proportion of data that belongs to the mixture sample n_m/n is determined by α and δ, and it is independent of n. However, the optimal $\epsilon(\delta_m)$ achievable involves n. How does the optimal proportion n_m/n change as δ_m varies within $(0, \delta)$, given fixed α and δ? How about the optimal ϵ, given a fixed n?

Figures 4 and 5 show how the optimal n_m/n and the optimal ϵ vary as δ_m changes, with $\alpha = 0.10$ and $\alpha = 0.90$, respectively. We see that unless α is large, the optimal ϵ is

usually achieved somewhere not far away from $\delta_m = \delta/2$, and the corresponding optimal n_m/n changes slowly near this area as well. If the guarantee desired on ϵ does not need to be precise, then when α is not large, a shortcut is to take $\delta_m = \delta/2$ and calculate the corresponding optimal n_m/n and optimal ϵ.

However, we can also find the precise optimal solution with respect to n_m and δ_m. As stated in Theorem 2, $\epsilon(\delta_m)$ is again a convex function in $\delta_m \in (0, \delta)$. We can solve the problem of minimizing $\epsilon(\delta_m)$ with respect to δ_m numerically, and the minimum ϵ we get is also the minimum of $\epsilon(n_m, \delta_m)$ optimized over n_m and δ_m together. In the following, we compute the numerical solution using a line search.

For fixed δ, how do the optimal proportion n_m/n, the corresponding optimal δ_m, and the optimal ϵ achievable change as the anomaly proportion α varies? In Figure 6 we see that as $\alpha \to 0$, which means we almost have no anomaly points in the mixture population, the optimal $n_m/n \to 0.5$, but the best epsilon achieved increases very rapidly. As $\alpha \to 1$, the mixture population is almost the same as the pure anomaly distribution. In this case, we want to allocate most of n to n_m, since focusing on the mixture distribution can almost directly give us a good threshold estimate. The trend that optimal $n_m/n \to 1$ and optimal ϵ guaranteed goes towards a small number align with this intuition. On the other hand, Figure 7 shows the optimal n_m/n and the corresponding δ_m as α varies. Here we observe that as $\alpha \to 0$, we are almost taking $\delta_m = \delta_0 = \delta/2$ and viewing estimating F_0 and F_m equally important. As $\alpha \to 1$, $\delta_m \to \delta$, estimating F_m becomes the most important consideration.

Given the budget constraint on n, confidence level $1 - \delta$, and anomaly proportion α, we can calculate the best ϵ achievable and the optimal n_m/n to obtain it. If we want to find the smallest n that could guarantee a certain ϵ value, we can solve the optimization problems for different values of n.

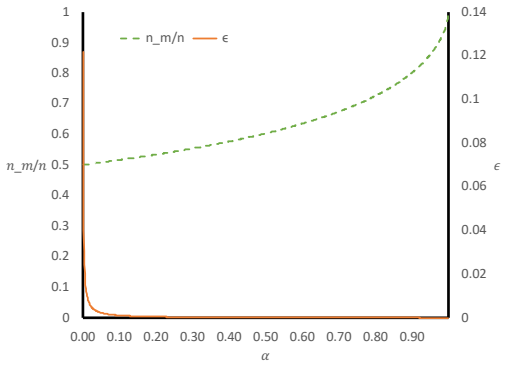

Figure 6: Optimal proportion of n allocated to mixture and the optimal ϵ guaranteed v.s. α, n = 1 000 000 000.

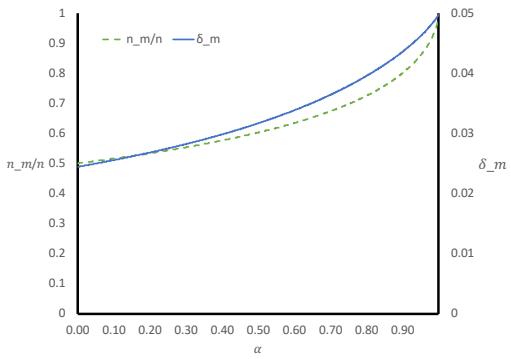

Figure 7: Optimal proportion of n allocated to mixture and corresponding δ_m v.s. α, n = 1 000 000 000.

7. Experiments on Threshold Estimation Given α

We performed experiments to answer four questions. Question Q1: how accurate is our estimate of $\hat{\tau}_q$ as a function of n and α? Question Q2: how loose are the bounds from Theorem 1? Question Q3: what are typical values of the false alarm rates for various settings of n and α on real data sets? Question Q4: how do these observed values change if we employ an overestimate $\alpha' > \alpha$?

All of our experiments employ the Isolation Forest anomaly detector of Liu et al. (2008), which has been demonstrated to be a state-of-the-art detector in recent empirical studies by Emmott et al. (2013).

Synthetic Data Experiments. To address Q1 and Q2, we run controlled experiments on synthetic data. The data points are generated from 9-dimensional normal distributions. The dimensions of the nominal distribution D_0 are independently distributed as $N(0, 1)$. The alien distribution is similar, but with probability 0.4, 3 of the 9 dimensions (chosen uniformly at random) are distributed as $N(3, 1)$ and with probability 0.6, 4 of the 9 dimensions (chosen uniformly at random) follow $N(3, 1)$. This ensures that the anomalies are not highly similar to each other and models the situation in which there are many different kinds of alien objects, not just a single alien class forming a tight cluster.

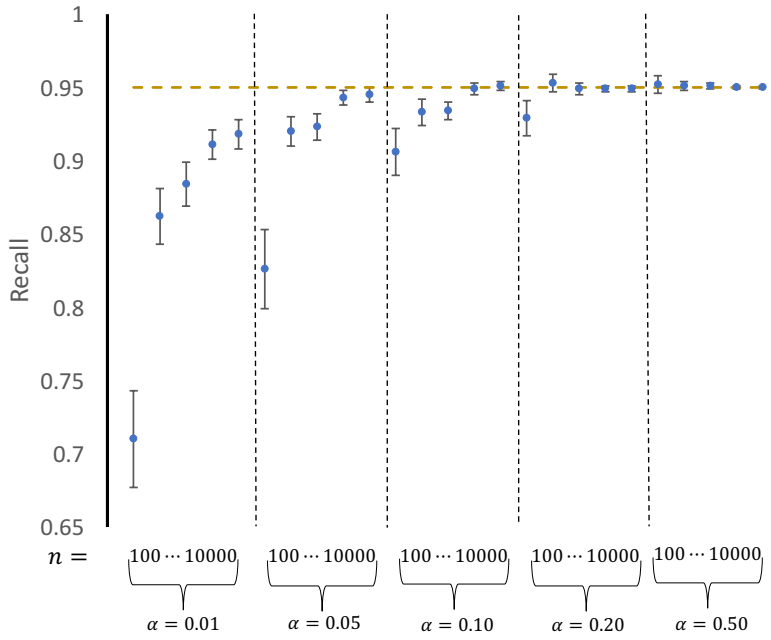

Figure 8: Comparison of recall achieved by $\hat{\tau}_q$ compared to oracle recall of 0.95. Error bars are 95% confidence intervals. Settings of n and α increase from left to right starting with $\alpha = 0.01$ and $n \in \{100, 500, 1\text{K}, 5\text{K}, 10\text{K}\}$ up to $\alpha = 0.5$ and $n = 10\text{K}$.

In each experiment, the nominal data set and the mixture data set are of the same size n, and the mixture data set contains a proportion α of anomaly points. We fixed the target

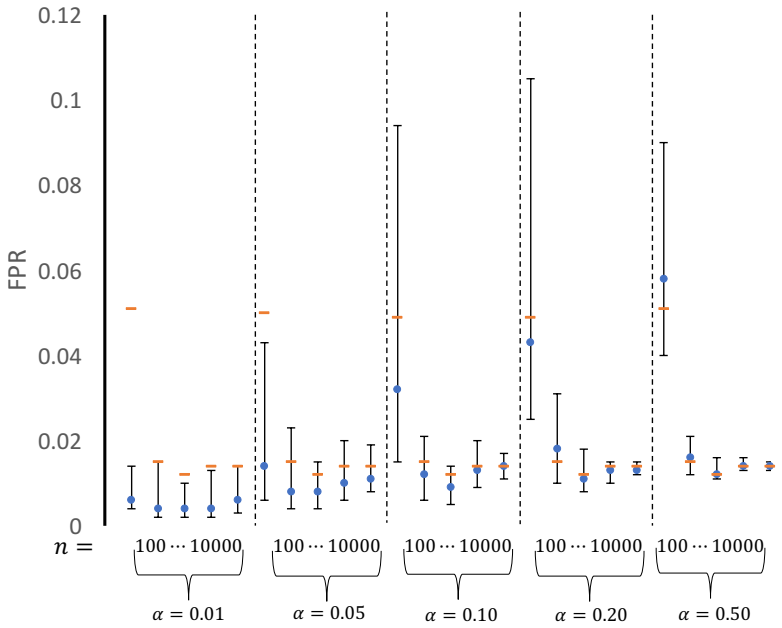

Figure 9: Comparison of oracle FPR to the FPR achieved by $\hat{\tau}_q$. Error bars span from the 25th to 75th percentile with the blue dot marking the median of the 100 trials. Orange markers indicate the oracle FPR. Settings of n and α increase from left to right starting with $\alpha = 0.01$ and $n \in \{100, 500, 1\text{K}, 5\text{K}, 10\text{K}\}$ up to $\alpha = 0.5$ and $n = 10\text{K}$.

quantile to be $q = 0.05$. The experiments are carried out for $n \in \{100, 500, 1\text{K}, 5\text{K}, 10\text{K}\}$ and $\alpha \in \{0.01, 0.05, 0.10, 0.20, 0.50\}$.

For testing, we create two large data sets G_0 and G_a, with G_0 being a pure nominal data set, G_a being a pure alien data set, and $|G_0| = |G_a| = 20K$. The Isolation Forest algorithm computes 1000 full depth isolation trees on the nominal data. Each tree is grown on a randomly-selected 20% subsample of the clean data points. We compute anomaly scores for the nominal points via out-of-bag estimates and anomaly scores for the mixture points, G_0, and G_a using the full isolation forest. This avoids the need for a separate validation set, and would be the correct procedure to follow in real applications. For each combination of n and α, we repeat the experiment 100 times. We measure the fraction of aliens detected (the "recall") and the fraction of nominal points declared to be alien (the "false positive rate") by applying the $\hat{\tau}_q$ estimate to threshold the anomaly scores in G_0 and G_a.

To assess the accuracy of our $\hat{\tau}_q$ estimates (Q1), we could compare them to the true values. However, this comparison is hard to interpret, because τ is expressed on the scale of anomaly scores, which are somewhat arbitrary. Instead, Figure 8 plots the recall achieved by $\hat{\tau}_q$. If $\hat{\tau}_q$ had been estimated perfectly, the recall would always be $1 - q = 0.95$. However, we see that the recall is often less than 0.95, which indicates that $\hat{\tau}_q$ is over-estimated, especially when n and α are small. This behavior is predicted by our theory, where we see that the sample size requirements grow inversely with α^2. For larger α and n, the recall guarantee is generally achieved. Figure 9 compares the false positive rate of the true oracle

τ_q to the false positive rate of the estimate $\hat{\tau}_q$. For each combination of α and n, we have 100 replications of the experiment and therefore 100 estimates $\hat{\tau}_a$ and 100 FPR rates. For each of these, the true FPR is computed using G_0. The error bars summarize the resulting 100 FPR values by the median and inter-quartile range. We see that for small n and α, the FPR can be quite different from the oracle rate, but for larger n and α, the estimates are very good.

To assess the looseness of the bounds (Q2), for each combination of n and α, we fix $\delta = 0.05$ and compute the value of η such that 95 of the 100 runs achieved a recall of at least $1 - \eta$ (thus η empirically achieves the $1 - \delta$ guarantee). We then compute $\epsilon = \eta - q$ and the corresponding required sample size n^* according to Theorem 1. Figure 10 shows a plot of n^* versus the actual n. The distance of these points from the $n^* = n$ diagonal line show that the theory is fairly loose, although it becomes tighter as n gets large.

Benchmark Data Experiments. To address our third and fourth questions, we performed experiments on six UCI multiclass data sets: Landsat, Opt.digits, pageb, Letter Recognition, Shuttle and Covertype. In addition to these, we provide results for the MNIST and Tiny ImageNet data sets. In each multiclass data set, we split the classes into two groups: nominal and alien. For Tiny ImageNet, we train a deep neural network classifier on 200 nominal classes and treat the remaining 800 as aliens. The nominal classes for UCI data sets are Landsat(1,7), OCR(1,3,4,5,7), pageb(1,5), Letter Recognition(1,3), Shuttle(1,4), and Covertype(1,2,3,7). The nominal classes for MNIST data set are (1,3,7). We generated nominal and mixture data sets for various values of α. The value of n for each data set is 1 600 for Landsat, 1 492 for OCR, 1 112 for pageb, 802 for Letter recognition, 8 777 for

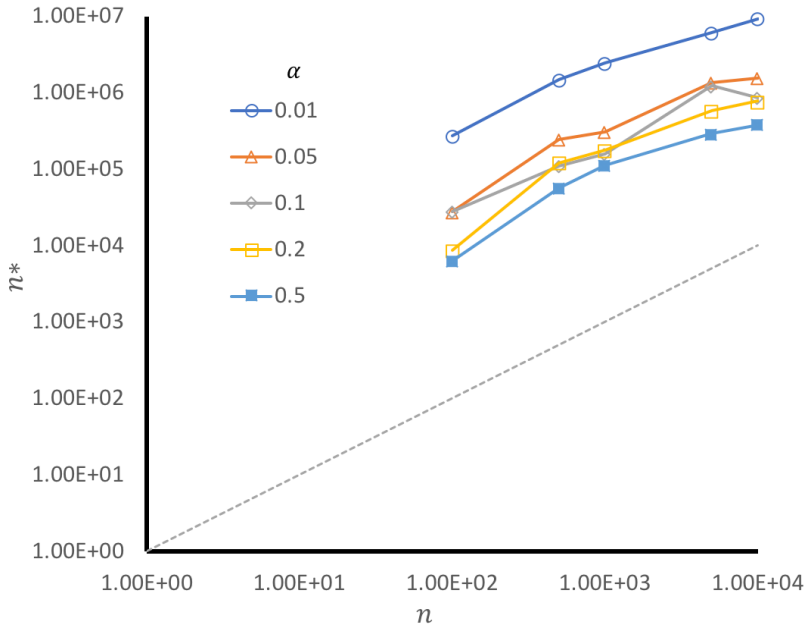

Figure 10: The log sample size n^* required by Theorem 1 in order to guarantee the actual observed recall versus the log actual sample size n.

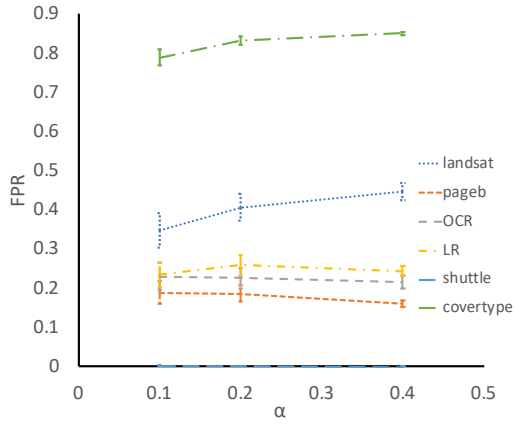

Figure 11: False positive rates on six UCI datasets as a function of α ($q = 0.05$).

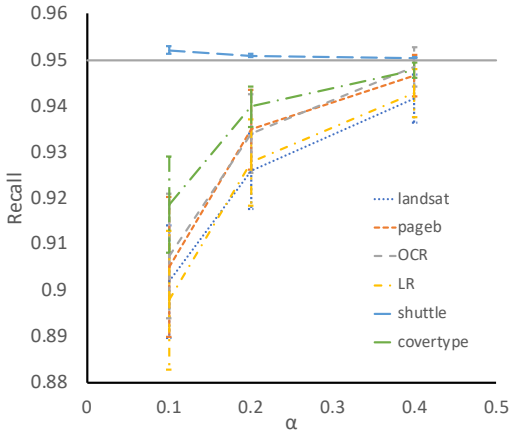

Figure 12: Recall rates on six UCI datasets as a function of α ($q = 0.05$)

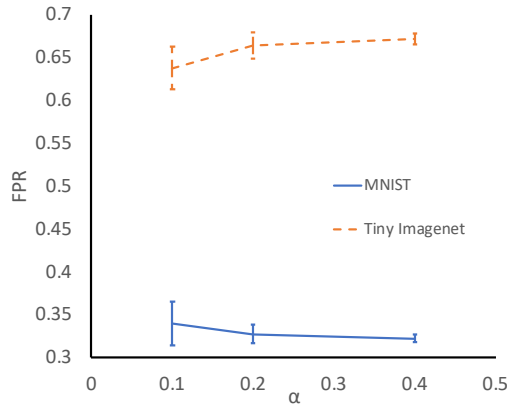

Figure 13: False positive rates on two image datasets as a function of α ($q = 0.05$).

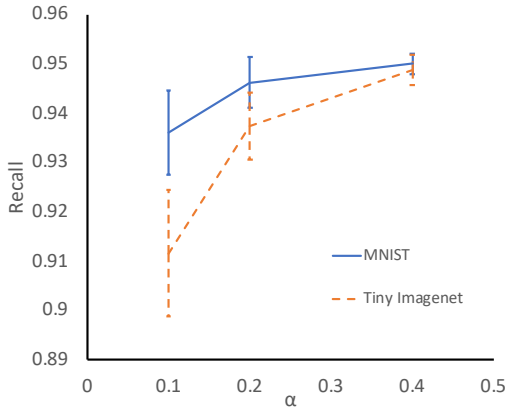

Figure 14: Recall rates on two image datasets as a function of α ($q = 0.05$).

Shuttle, 10 000 for Covertype, 10 000 for MNIST, and 5 263 for Tiny ImageNet. Because we cannot create data sets with large n, we cannot measure the true value of τ_q.

After computing the anomaly scores for both nominal and mixture data sets, we applied Algorithm 1 within a 10-fold cross validation. We divide the mixture data points at random into 10 groups. For each fold, we estimate $\hat{F}_a$ and $\hat{\tau}_a$ from 9 of the 10 groups and then score the mixture points in the held-out fold according to $\hat{\tau}_a$. In all other respects, the experimental protocol is the same as for the synthetic data. For Tiny ImageNet, the anomaly scores are obtained by applying the baseline method of Hendrycks and Gimpel (2017).

To answer Q3, Figures 11 and 13 plot the false positive rate as a function of α for the UCI and vision data sets, respectively. We see that the FPR ranges from 0.06% to 85.05%

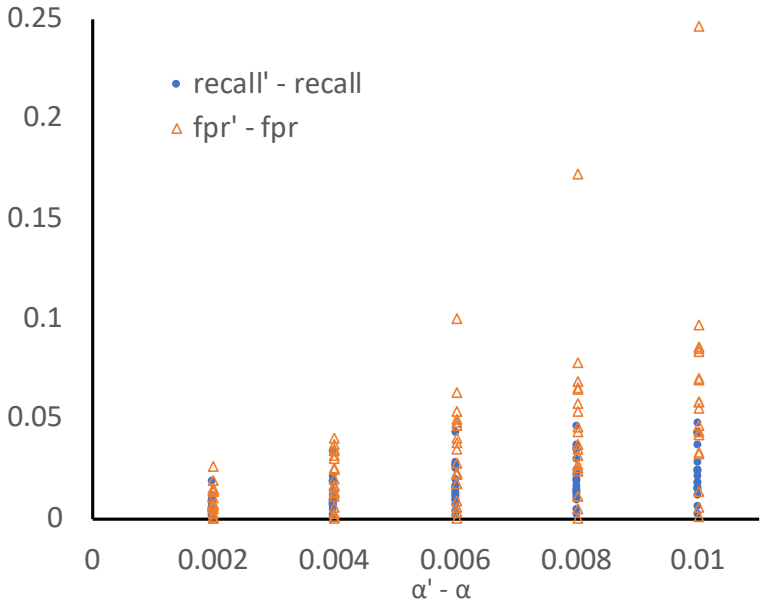

Figure 15: Change in recall and false positive rate as a function of $\alpha' - \alpha$ for six UCI datasets; $\alpha \in \{0.1, 0.2, 0.4\}$

on UCI depending on the data set and the level of α. The vision data sets have higher FPR, especially Tiny ImageNet, which has a large number of alien classes that are not distinguished well by the anomaly detector. The FPR depends primarily on the problem domain and the performance of anomaly detector, because the key issue is how well the anomaly detector distinguishes between nominal and alien examples. For some data sets, the FPR can be improved by applying an anomaly detector better-suited to the problem. While for other data sets where the anomalies and nominals are hard to tell apart in nature, it may not be feasible to improve FPR without sacrificing recall.

Figures 12 and 14 plot the recall as a function of α for the UCI and vision data sets. We set $q = 0.05$ in these experiments. Theorem 1 only guarantees a recall of $1 - q - \epsilon$, where ϵ depends on n. Hence, it is nice to see that for two of the domains (Shuttle and Covertype) in UCI data sets and for both vision data sets the recall is very close to $1 - q = 0.95$. These are the domains with the largest values of n. The value of α has a bigger impact on recall than it does on FPR. This is because the effective number of alien training examples is αn, which can be very small for some data sets when $\alpha = 0.1$. The recall generally improves as α increases. In some applications, it may be possible to enrich S_m so that α is larger on the training set to take advantage of this phenomenon. It is interesting to note that once $\hat{\tau}_a$ has been computed, it can be applied to test data sets having different (or unknown) values of α. If the mixture data set S_m cannot be enriched to increase α, then when α is very small, S_m needs to be very large.

To answer Q4 regarding the impact of using an incorrect value $\alpha' > \alpha$, we repeated these experiments on UCI data sets with $\alpha' = \alpha + \xi$, for $\xi \in \{0.002, 0.004, 0.006, 0.008, 0.010\}$. Figure 15 plots the change in false positive rate and recall as a function of $\alpha' - \alpha$. Two points

are plotted for each combination of α' and data set, the change in recall and the change in FPR. We observe that the recall increases slightly (in the range from 0.001 to 0.047). However, the false positive rate increases by much larger amounts (from 0 to 0.246). This demonstrates that it is very important to determine the value of α accurately to control the false positive rate.

8. Threshold Estimation via α Estimation

In many practical applications, the exact value or a reasonable upper bound of α is not available. In this section we combine certain mixture proportion estimators with our algorithm to explore a possible solution for estimating the threshold in practice.

Multiple methods in the literature exist for estimating α. One problem that is jointly recognized by multiple studies (e.g., Scott, 2015; Patra and Sen, 2016; Ramaswamy et al., 2016) is that the exact α is not identifiable in the general setting (i.e., without making additional assumptions about the relationship of F_0 and F_a). To see why, suppose that

$$F_m(x) = (1-\alpha)F_0(x) + \alpha F_a(x)$$

and we know the exact form of $F_m(x)$ and $F_0(x)$. Then for any $\alpha' > \alpha$, $F_m(x)$ can rewritten as

$$F_m(x) = (1-\alpha')F_0(x) + \alpha' F'_a(x),$$

where

$$F'_a(x) = \frac{(\alpha'-\alpha)F_0 + \alpha F_a(x)}{\alpha'}.$$

This implies that without additional assumptions, any $\alpha' \geq \alpha$ will provide us a legal way to decompose F_m. In the extreme case, taking $\alpha' = 1$ will just make F'_a equal to F_m. Thus, in the literature, most studies either enforce a separability condition between F_0 and F_a (Scott, 2015; Ramaswamy et al., 2016) or focus on estimating

$$\alpha_0 := \inf\{\gamma \in (0,1] : \frac{F_m - (1-\gamma)F_0}{\gamma} \text{ is a CDF}\}, \tag{3}$$

which means we attribute as large a proportion of F_m as possible to F_0, as long as a legal CDF for the anomalous part F_a can still be reconstructed (Patra and Sen, 2016).

In this section, we consider nine different estimators for α_0 and demonstrate their performance on synthetic data sets. Additionally, we compute the resulting recall and false positive rate (FPR) from five of the estimators on selected UCI benchmark data sets when deployed to estimate τ_α for open category detection.

The first four estimators are based on Ramaswamy et al. (2016), where the authors embed both the empirical clean and mixture distributions into a reproducing kernel Hilbert space (RKHS) and then estimate α_0 by finding the smallest value of γ such that the estimated $\hat{P}_a$ is the embedding of a probability distribution. This is the same intuition as Equation 3, but instead of requiring $\hat{F}_a$ to be a valid CDF, they require $\hat{P}_a$ to be a valid probability distribution. They do not actually require $\hat{P}_a$ to be a valid embedding, but instead they threshold the distance between $\hat{P}_a$ and a valid embedding. The paper proposes two methods, "KM1" and "KM2", where KM1 thresholds the distance directly and KM2 thresholds it based on its gradient.

In our setting, we can apply these algorithms in two separate ways. First, they can be applied directly to the samples S_0 and S_m. We refer to the corresponding estimates as α_1 and α_2. The second way is to first compute anomaly scores (e.g., via an isolation forest), and then embed those scores into the RKHS. We refer to the resulting estimators as $\alpha_{1\text{score}}$ and $\alpha_{2\text{score}}$.

Among the next four estimators, one is extended by Lin and Long (2020) from an estimator introduced by Patra and Sen (2016), and the other three come from two sources and are given a good summary by Lin and Long (2020). The main framework of Lin and Long (2020) trains a flexible classifier, such as a random forest, to discriminate between the nominal sample S_0 and the mixture sample S_m. The CDFs G_0 and G_m of the fitted probability scores of this classifier satisfy the relationship that $G_m = (1-\alpha)G_0 + \alpha G_a$. Lin and Long compare four mixture proportion estimators, namely "C-PS", "C-ROC", "ROC" and "SPY". Here "C-PS" is the estimator extended from the original version by Patra and Sen (2016) to the case when F_0 is not known, and the letter "C" indicates that it is based on the fitted classifier. "C-ROC" and "ROC" are both based on the method proposed by Scott (2015) that makes use of the classifier's ROC curve to estimate α, by taking 1 minus the minimum slope between the point $(1, 1)$ and any other point on the curve. The difference is that "C-ROC" is based on a random forest classifier and "ROC" is based on kernel logistic regression.

The "SPY" method was proposed by Liu et al. (2002). In the implementation, Lin and Long (2020) fit a random forest to separate S_0 from S_m. To determine a probability threshold for labeling points in S_m as nominal versus anomalous, the "SPY" method first inserts a modest number of clean points from S_0 (the "spies") into S_m. By analyzing the class membership probabilities assigned by random forest to the spies, the method selects a threshold and labels all points from the original S_m with assigned probabilities lower than the threshold to be "reliable negatives". These are assigned labels of 0, the points in the original S_0 are assigned labels of 1, and another random forest is built to tell these "reliable negatives" apart from points in the original S_0. In the last step, the latter random forest is applied on the original S_m to obtain the predicted labels, and the proportion of points predicted to be anomalies is taken as an estimate for α.

In addition to the eight methods already introduced, we propose a new method inspired by Patra and Sen (2016), but based on bootstrap reconstruction of the mixture sets to select the best value of γ. The procedure is described in Algorithm 2.

The intuition behind the original method by Patra and Sen (2016) is, when $\gamma < \alpha_0$, we are attributing too much of the F_m to F_0, and $F_a = \frac{F_m - (1-\gamma)F_0}{\gamma}$ will not be a legal CDF. As a result, the $\hat{F}_a^\gamma$ we compute from $\hat{F}_m$ and $\hat{F}_0$ will also tend to be far away from a legal CDF—specifically, it will be far away from the CDF $\check{F}_a^\gamma$ computed by applying isotonic regression to $\hat{F}_a^\gamma$. Patra and Sen define the distance of two functions g and h on the real line as $d_m(g, h) = \int \{g(x) - h(x)\}^2 dF_m(x)$. Hence, the distance between the original F_m and the reconstructed version based on $\check{F}_a^\gamma$ can be written as $d_m(F_m, \gamma\check{F}_a^\gamma + (1-\gamma)F_0) = \gamma d_m(\hat{F}_a^\gamma, \check{F}_a^\gamma)$. Patra and Sen (2016) prove that $\gamma d_m(\hat{F}_a^\gamma, \check{F}_a^\gamma)$ converges almost surely to 0 when $\gamma \geq \alpha_0$ and to a positive quantity when $\gamma < \alpha_0$ as the sample size increases. Hence, we can select as our estimator of α_0 the value of γ that makes this distance small. Because of random fluctuations due to sampling, we do not expect that $d_m(\hat{F}_a^\gamma, \check{F}_a^\gamma)$ will be equal to

Algorithm 2 Bootstrap-Based Mixture Proportion Estimation

for $\gamma = 0.005, 0.010, \ldots, 1$ **do**
  Compute $d(\gamma) = \gamma d_m(\hat{F}_a^\gamma, \check{F}_a^\gamma)$.
  for $b = 1, \ldots, B$ **do**
    Get bootstrap sample S_m^b by sampling nominal points with replacement from S_0 and anomalies from $\check{F}_a^\gamma$
    Compute resampled estimates $\hat{F}_{a,b}^\gamma$ and $\check{F}_{a,b}^\gamma$
    Compute the resampled distance $d_b(\gamma) = \gamma d_{m,b}(\hat{F}_{a,b}^\gamma, \check{F}_{a,b}^\gamma)$
  end for
  Compute the 25% and 75% quantiles from $d_1(\gamma), d_2(\gamma), \ldots, d_B(\gamma)$ as $d^{0.25}(\gamma)$ and $d^{0.75}(\gamma)$
  if $d(\gamma) \in [d^{0.25}(\gamma), d^{0.75}(\gamma)]$ **then**
    break and output $\hat{\alpha} = \gamma$
  end if
  if $d(\gamma) \notin [d^{0.25}(\gamma), d^{0.75}(\gamma)]$ and $\gamma = 1$ **then**
    output $\hat{\alpha} = 1$
  end if
end for

zero. Hence, the central question becomes determining what distance we would expect to have due to random fluctuations when $\gamma = \alpha_0$?

We propose to answer this question via bootstrap resampling. Specifically, if $\gamma = \alpha_0$, then if we resample the anomalous points S_a according to $\check{F}_a^\gamma$, mix them with clean points from S_0 according to mixing proportion γ and re-estimate $\hat{F}_a^\gamma$ and $\check{F}_a^\gamma$, we should obtain a bootstrap replicate of the distance $d_m(\hat{F}_a^\gamma, \check{F}_a^\gamma)$. Repeating this many times will allow us to assess the typical distance due to sampling fluctuations. If the bootstrap replicate distances are similar to the $d_m(\hat{F}_a^\gamma, \check{F}_a^\gamma)$, then we can adopt γ as our estimate of α_0.

Specifically, from the bootstrap replicates, we compute the 25% and 75% quantiles of the distribution of $d_m(\hat{F}_a^\gamma, \check{F}_a^\gamma)$ and select as our estimate $\hat{\alpha}_0$ the smallest value of γ such that the computed $d_m(\hat{F}_a^\gamma, \check{F}_a^\gamma)$ falls within this inter-quartile range of the bootstrap replicates.

The choice of the 25% and 75% quantiles is entirely heuristic. Selecting a higher upper bound will tend to produce a smaller estimate for α_0.

We run experiments on synthetic data sets to compare the performance of nine α_0 estimators and summarize the performance in Figure 16. In these experiments, the generating process of synthetic data sets follows the one described in Section 7, and experiments for each setting are repeated 100 times. Due to the construction method of the data sets, we can assume that the nominal and anomalous distributions approximately satisfy the separability requirements, and thus $\alpha_0 \approx \alpha$.

We observe that when both α_0 and the sample size are small, both α_1 and α_2 estimators severely overestimate α, but the performance improves as α increases and as the sample size increases. The anomaly-score-based methods $\alpha_{1\text{score}}$ and $\alpha_{2\text{score}}$ appear slightly better than α_1 and α_2 which suggests that converting the raw data to anomaly scores first reduces the large positive bias. All estimators, except the "SPY" method, show a trend toward the true α value as the sample size increases. However, the C-ROC and ROC methods

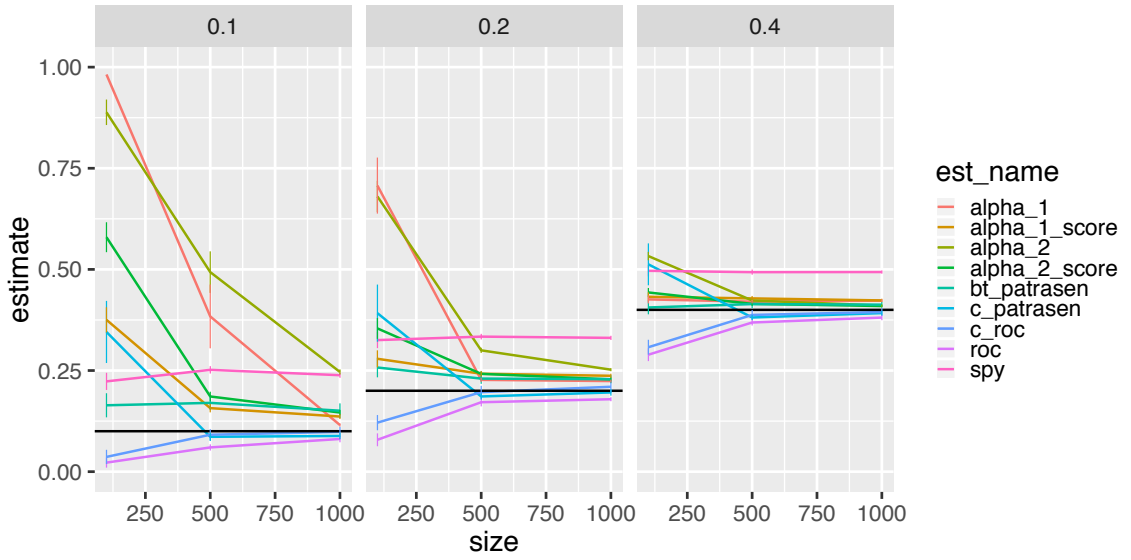

Figure 16: Average estimate and error bars of 9 different estimators for α_0, with each plot corresponding to one α_0 value in $(0.10, 0.20, 0.40)$, and sample sizes in $(100, 500, 1000)$. The black horizontal line shows the true value of α_0.

tend to underestimate, and α_1, $\alpha_{1\text{score}}$, α_2, $\alpha_{2\text{score}}$, and our method (bt-patrasen) tend to overestimate.

Next, we apply five of these methods to estimate τ_q by first estimating α and then applying our Algorithm 1. We did not include the methods based on Ramaswamy et al. (2016), namely α_1, α_2, $\alpha_{1\text{score}}$ and $\alpha_{2\text{score}}$, because they employ quadratic programming as part of the procedure and thus take relatively large computing resources. Based on the insight from the classifier-based methods for estimating α, we perform these experiments using two different configurations: one based on iForest anomaly scores and one based on class probabilities produced by a flexible classifier. The configurations for these experiments are shown in Tables 1 and 2.

In these experiments, the classes we treat as known classes from each data set are LR (1,3), pageb (1,5), OCR(1,3,4,5,7), landsat(1,7), shuttle(1,4), and covertype (1,2,3,7). The data set size $n = |S_0| = |S_m|$ is LR (802), pageb (1 112), OCR (1 492), landsat (1 600), shuttle (8 777), and covertype (10 000).

Figure 17 shows the iForest-based results and Figure 18 shows the classifier-based results. The data set names on x-axis in both figures are ordered by size. We see that in both configurations, the performance of the various proportion estimators are similar, except that the SPY method consistently overestimates α and consequently achieves high false positive rates. In general, as the sample size of the data sets increases, the estimate of the alien fraction (α) gets closer to its true value, recall gets closer to $1 - q$, and the false positive rate gets closer to that achieved when under the true value of α. These results show that combining anomaly proportion estimation and our algorithm of setting threshold is an effective approach for real-world data.

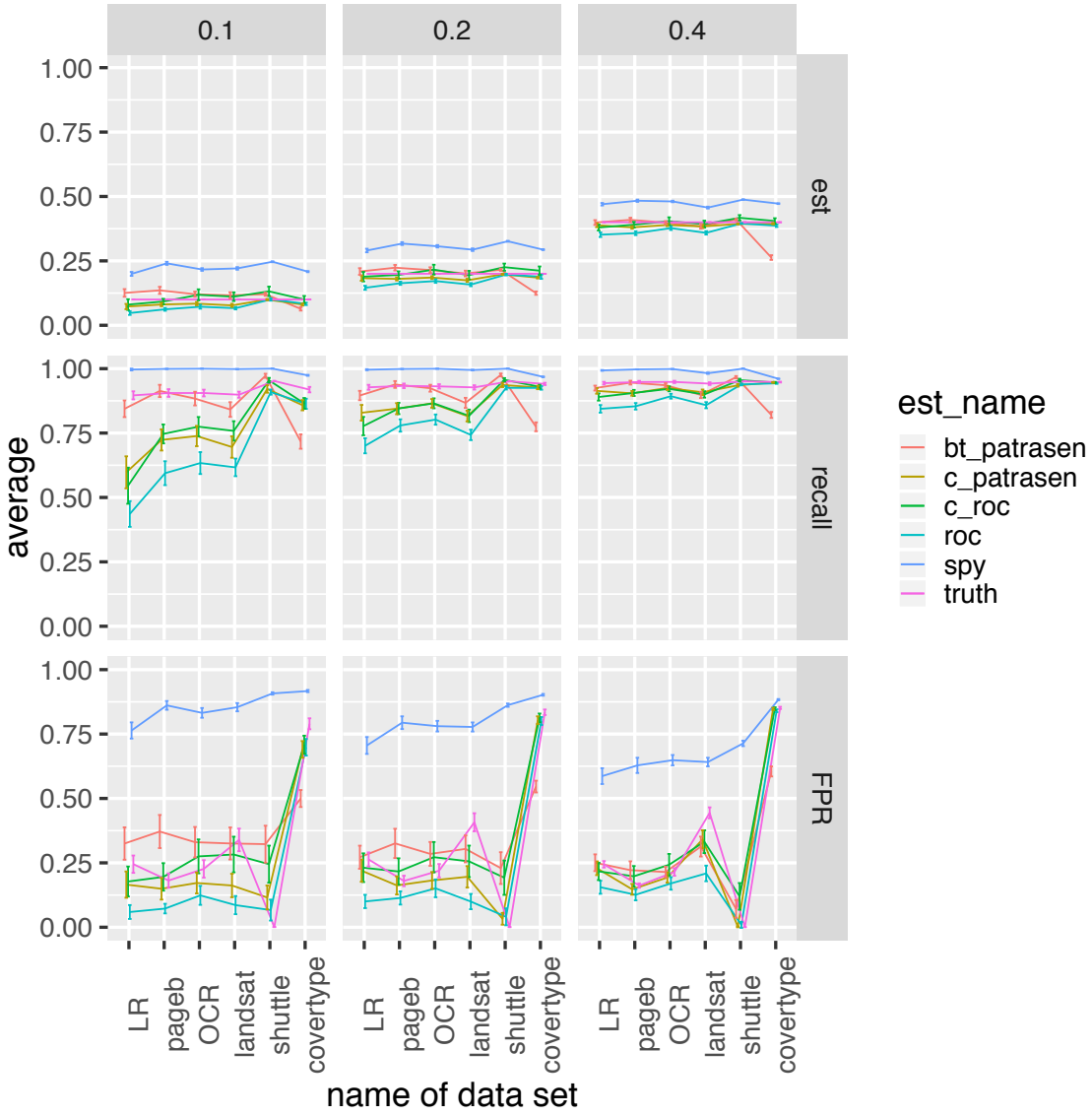

Figure 17: With iForest anomaly scores: the average estimate and error bars on proportion estimation, recall, and FPR achieved by 5 different estimators for α_0, with α_0 value in $(0.10, 0.20, 0.40)$, and over 6 different UCI data sets. The "truth" here corresponds to the values obtained using the true $\alpha(q = 0.05)$.

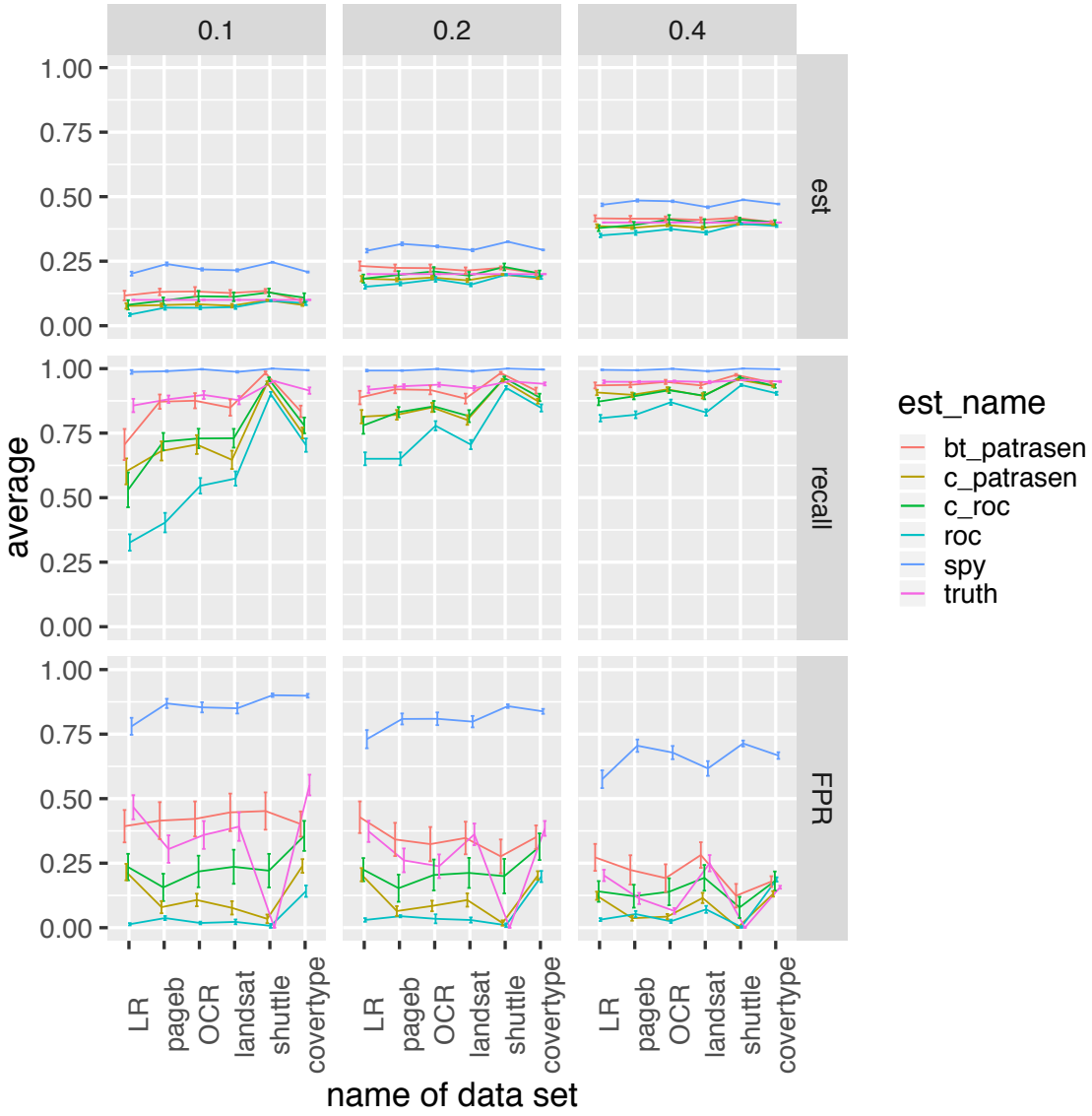

Figure 18: With classifier-provided class probabilities: the average estimate and error bars on proportion estimation, recall, and FPR achieved by 5 different estimators for α_0, with α_0 value in $(0.10, 0.20, 0.40)$, and over 6 different UCI data sets. The "truth" here corresponds to the values obtained using the true $\alpha(q = 0.05)$.

estimator	proportion estimate	threshold estimate
bt-patrasen	iForest scores	iForest scores
C-PS	random forest scores	iForest scores
C-ROC	random forest scores	iForest scores
ROC	kernel logistic regression scores	iForest scores
SPY	random forest scores	iForest scores
truth	true α value	iForest scores

Table 1: Experiment configuration using iForest scores

estimator	proportion estimate	threshold estimate
bt-patrasen	random forest scores	random forest scores
C-PS	random forest scores	random forest scores
C-ROC	random forest scores	random forest scores
ROC	kernel logistic regression scores	kernel logistic regression scores
SPY	random forest scores	random forest scores
truth	true α value	random forest scores

Table 2: Experiment configuration using classifier scores

Our method, bt_patrasen, is one of the best-performing methods, except on covertype with iForest anomaly scores. The iForest anomaly detector appears to perform very badly on the covertype data, and all estimators—and the true α value—give poor false positive rates. However, all of the other estimators still produce good estimates of α in this case, but bt_patrasen gives a serious underestimate. This indicates that the random forest and kernel logistic regression classifiers provide better scores for the covertype problem for setting the threshold τ. The tradeoff achievable between recall and false positive rate is determined by the nature of the problem and the performance of the anomaly detector and classifier.

For dealing with problems in practice, we recommend first computing three estimates of α from bt_patrasen, c_patrasen and c_roc. If these estimates are similar, we recommend using bt_patrasen. Otherwise use c_roc. If it is known from background knowledge that a certain anomaly detector or classifier in general works well for the problem, then we suggest using its scores for estimating α and especially for estimating τ_q.

9. Discussion

We have taken a step toward open category detection with guarantees by providing a PAC-style guarantee on the probability of detecting $1 - \eta$ of the aliens on the test data. This is the first such guarantee under any similarly general conditions. We have shown that this guarantee is satisfied in our experiments, although the guarantee is somewhat loose, especially on small training sets. Obtaining a guarantee requires more data than standard PAC guarantees on expected prediction accuracy. This is because we must estimate the q quantile of the alien anomaly score distribution, where q is typically quite small. Nonethe-

less, our experiments show that our algorithm gives good recall performance and non-trivial false alarm rates on data sets of reasonable size.

It is important to note that the very formulation of a PAC-style guarantee on the probability of detecting aliens requires assuming that the aliens are drawn from a well-defined distribution D_a. While this is appropriate in some applications, such as the insect survey application described in the introduction, it is not appropriate for adversarial settings. In such settings, a PAC-style guarantee does not make sense, and some other form of safety guarantee needs to be formulated.

To obtain the guarantee, we employ two training data sets: a clean data set that contains no aliens and an (unlabeled) contaminated data set that contains a known fraction α of aliens. As we have seen in the experiments, if we don't know the exact value of the alien proportion α in the mixture distribution, then having a tight upper bound for α is important; otherwise the false positive rate grows rapidly. To the best of our knowledge, no method is known that can provide a non-trivial upper bound on α within the setup of this chapter. Blanchard et al. (2010) show that if the clean data distribution D_0 is weakly diffuse, then there is no distribution-free non-trivial upper confidence bound for α_0. (It is possible to obtain a bound if the support of D_0 is finite.) As a consequence, we cannot get a distribution-free non-trivial upper confidence bound in our setting. Sanderson and Scott (2014) and Patra and Sen (2016) propose two different consistent estimators, but these do not give an upper bound estimate with explicit finite sample guarantees. Despite that, Corollary 2 is effective if an expert has domain knowledge on what a sound upper bound would be.

Our guarantee requires more data as α becomes small. Fortunately, when α is small, it may be possible in some applications to afford lower recall rates, since the frequency of aliens will be smaller. However, in safety-critical applications where a single undetected alien poses a serious threat, there is little recourse other than to collect more data or allow for higher false positive rates.

To deal with the problem estimating the threshold τ_q when an upper bound on α is not directly available, we evaluated five α estimators with Algorithm 1 on six UCI data sets. With sufficient training data, these estimators produced accurate estimates for α, high anomaly detection rates, and low false positive rates. Hence, the combination of these estimators for α and our method of setting τ_q gives a practical approach to obtaining good performance on real-world problems.

Of course the performance of these methods, especially on FPR, is affected by the performance of the anomaly detector and classifier (when using the classifier-based α estimation methods). In an application setting, preliminary experiments should be run to select a good classifier and a good anomaly detector to employ for estimating α and selecting τ_q.

Acknowledgments

This research was supported by the National Science Foundation under grant number 1514550, DARPA contract HR001120C0022 (to Raytheon BBN Technologies), Air Force contract FA8750-19-C-0092 (to Galois, Inc.), and a gift from Huawei, Inc. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the

author(s) and do not necessarily reflect the views of the sponsors. The authors also thank Tadesse ZeMicheal for providing his implementation of iForest.

Appendix A. Proofs

A1. Proof for Theorem 1

Suppose there are n random variables which are i.i.d. from the distribution with CDF F and let $\hat{F}_n$ be the empirical CDF calculated from this sample. Then Massart (1990) shows that

$$P(\sqrt{n}\sup_x |\hat{F}_n(x) - F(x)| > \lambda) \leq 2\exp(-2\lambda^2) \tag{4}$$

holds without any restriction on λ. Making use of this, and assuming we use the same sample size n for both the mixture data set and the clean data set, for any $\epsilon \in (0, 1-q)$, we seek to determine how large n needs to be in order to guarantee that with probability at least $1-\delta$ our quantile estimate $\hat{\tau}_q$ satisfies $F_a(\hat{\tau}_q) \leq q + \epsilon$. To achieve this, we want to have

$$P(\sup_x |\hat{F}_a(x) - F_a(x)| > \epsilon) \leq \delta.$$

We have

$$\begin{aligned}
& P(\sup_x |\hat{F}_a(x) - F_a(x)| > \epsilon) \\
= \; & P(\sup_x |\frac{\hat{F}_m(x) - (1-\alpha)\hat{F}_0(x)}{\alpha} - \\
& \frac{F_m(x) - (1-\alpha)F_0(x)}{\alpha}| > \epsilon) \\
= \; & P(\sup_x |\frac{1}{\alpha}(\hat{F}_m(x) - F_m(x)) - \\
& \frac{1-\alpha}{\alpha}(\hat{F}_0(x) - F_0(x))| > \epsilon) \\
\leq \; & P((\frac{1}{\alpha}\sup_x |\hat{F}_m(x) - F_m(x)| + \\
& \frac{1-\alpha}{\alpha}\sup_x |\hat{F}_0(x) - F_0(x)|) > \epsilon) \\
\leq \; & P(\{\frac{1}{\alpha}\sup_x |\hat{F}_m(x) - F_m(x)| > \frac{1}{2-\alpha}\epsilon\} \\
& \cup \{\frac{1-\alpha}{\alpha}\sup_x |\hat{F}_0(x) - F_0(x)| > \frac{1-\alpha}{2-\alpha}\epsilon\}) \\
= \; & P(\{\sup_x |\hat{F}_m(x) - F_m(x)| > \frac{\alpha}{2-\alpha}\epsilon\} \\
& \cup \{\sup_x |\hat{F}_0(x) - F_0(x)| > \frac{\alpha}{2-\alpha}\epsilon\}).
\end{aligned}$$

Making use of (4), when

$$n > \frac{1}{2}\ln\frac{2}{1-\sqrt{1-\delta}}(\frac{1}{\epsilon})^2(\frac{2-\alpha}{\alpha})^2,$$

we will have

$$P(\sup_x |\hat{F}_m(x) - F_m(x)| > \frac{\alpha}{2-\alpha}\epsilon) \leq 1 - \sqrt{1-\delta},$$
$$P(\sup_x |\hat{F}_0(x) - F_0(x)| > \frac{\alpha}{2-\alpha}\epsilon) \leq 1 - \sqrt{1-\delta}.$$

In this case we will have

$$\begin{aligned} & P(\sup_x |\hat{F}_a(x) - F_a(x)| > \epsilon) \\ \leq \; & 1 - P(\{\sup_x |\hat{F}_m(x) - F_m(x)| \leq \frac{\alpha}{2-\alpha}\epsilon\} \\ & \cap \{\sup_x |\hat{F}_0(x) - F_0(x)| \leq \frac{\alpha}{2-\alpha}\epsilon\}) \\ \leq \; & 1 - (1 - 1 + \sqrt{1-\delta})^2 \\ = \; & \delta. \end{aligned}$$

Now we have with probability at least $1-\delta$,

$$|\hat{F}_a(x) - F_a(x)| \leq \epsilon, \quad \forall x \in R.$$

If this inequality holds, then for any value $\hat{\tau}_q$ such that $\hat{F}_a(\hat{\tau}_q) \leq q$, we have

$$F_a(\hat{\tau}_q) \leq \hat{F}_a(\hat{\tau}_q) + \epsilon \leq q + \epsilon.$$

So we have with probability at least $1-\delta$, any $\hat{\tau}_q$ satisfying $\hat{F}_a(\hat{\tau}_q) \leq q$ will satisfy $F_a(\hat{\tau}_q) \leq q + \epsilon$. ■

A2. Proof for Corollary 1

Through Bonferroni correction, we split the probability tolerance δ equally into getting guarantee for recall and estimating the false positive rate resulted from threshold $\hat{\tau}_q$.

Following the reasoning of Theorem 1, we can get that as long as the sample size $|S_m| = |S_0| = n$ satisfy

$$n > \frac{1}{2} \ln \frac{2}{1 - \sqrt{1 - \delta/2}} (\frac{1}{\epsilon})^2 (\frac{2-\alpha}{\alpha})^2, \tag{5}$$

we can have that with probability at least $1 - \delta/2$, $F_a(\hat{\tau}_q) \leq q + \epsilon$.

Next, We want to get the minimum value of ϵ_0, such that

$$P(|\hat{F}_0(\hat{\tau}_q) - F_0(\hat{\tau}_q)| > \epsilon_0) \leq \delta/2.$$

Making use of equation (4) again, for the CDF of anomaly scores from clean data, we have

$$P(\sqrt{n} \sup_x |\hat{F}_0(x) - F_0(x)| > \lambda) \leq 2\exp(-2\lambda^2)$$

holds without any restriction on λ. Setting $2\exp(-2\lambda^2) \leq \delta/2$, we get $\lambda \geq \sqrt{\frac{1}{2}\ln\frac{4}{\delta}}$ and thus the minimum value of ϵ_0 is $\frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\frac{4}{\delta}}$. Now we have with probability at least $1 - \delta/2$,

$$F_0(\hat{\tau}_q) \geq \hat{F}_0(\hat{\tau}_q) - \epsilon_0.$$

Thus we have that with probability at least $1-\delta$,

$$1 - F_a(\hat{\tau}_q) \geq 1 - (q+\epsilon) \text{ and } 1 - F_0(\hat{\tau}_q) \leq 1 - \hat{F}_0(\hat{\tau}_q) + \epsilon_0.$$

On the other hand, from inequality (5), we have $\epsilon > \frac{1}{\sqrt{n}} \frac{2-\alpha}{\alpha} \sqrt{\frac{1}{2} \ln \frac{2}{1-\sqrt{1-\delta/2}}}$. Since $\alpha \in (0,1)$ and $\delta \in (0,1)$, we have

$$\epsilon_0 = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{2} \ln \frac{4}{\delta}} < \frac{1}{\sqrt{n}} \frac{2-\alpha}{\alpha} \sqrt{\frac{1}{2} \ln \frac{2}{1-\sqrt{1-\delta/2}}} < \epsilon.$$

■

A3. Proof for Corollary 2

If $\alpha' \geq \alpha$, and if we write

$$F'_a(x) = \frac{F_m(x) - (1-\alpha')F_0(x)}{\alpha'},$$

then F'_a is still a legal CDF, because

$$F'_a(-\infty) = 0, \quad F'_a(\infty) = 1,$$

and it is easy to show that F'_a is monotonically nondecreasing.

But

$$F'_a(x) - F_a(x) = \frac{(\alpha - \alpha')(F_m(x) - F_0(x))}{\alpha\alpha'} \geq 0, \forall x \in \Re,$$

and because of this, if we let $\hat{\tau}'_q$ denote the threshold we get from using α', we will have $F_a(\hat{\tau}'_q) \leq F'_a(\hat{\tau}'_q)$. By the proof of previous theorem, we know that when $n > \frac{1}{2} \ln \frac{2}{1-\sqrt{1-\delta}} (\frac{1}{\epsilon})^2 (\frac{2-\alpha'}{\alpha'})^2$, we have with probability at least $1-\delta$, $F'_a(\hat{\tau}'_q) \leq q + \epsilon$, and thus we have $F_a(\hat{\tau}'_q) \leq q + \epsilon$. ■

A4. Proof for Theorem 2

As mentioned before, our objective function is the guaranteed ϵ in terms of n_m and δ_m, given fixed n, δ and α:

$$\epsilon(n_m, \delta_m) = \frac{1}{\alpha} \frac{1}{\sqrt{n_m}} \sqrt{\frac{1}{2} \ln \frac{2}{\delta_m}} + \frac{1-\alpha}{\alpha} \frac{1}{\sqrt{n-n_m}} \sqrt{\frac{1}{2} \ln \frac{2}{\delta - \delta_m}},$$

where

$$0 < n_m < n, \qquad 0 < \delta_m < \delta.$$

To prove its convexity, we can write $\epsilon(n_m, \delta_m)$ into two parts:

$$\begin{aligned} g(n_m, \delta_m) &= \frac{1}{\alpha} \frac{1}{\sqrt{n_m}} \sqrt{\frac{1}{2} \ln \frac{2}{\delta_m}}, \\ h(n_m, \delta_m) &= \frac{1-\alpha}{\alpha} \frac{1}{\sqrt{n-n_m}} \sqrt{\frac{1}{2} \ln \frac{2}{\delta - \delta_m}}. \end{aligned}$$

Now

$$\epsilon(n_m, \delta_m) = g(n_m, \delta_m) + h(n_m, \delta_m).$$

First look at the components of the Hessian matrix of $g(n_m, \delta_m)$:

$$\begin{aligned}
\frac{\partial^2 g}{\partial n_m^2} &= \frac{3}{4}\frac{1}{\alpha} n_m^{-\frac{5}{2}} (\frac{1}{2}\ln\frac{2}{\delta_m})^{\frac{1}{2}}, \\
\frac{\partial^2 g}{\partial n_m \partial \delta_m} &= \frac{1}{8}\frac{1}{\alpha} n_m^{-\frac{3}{2}} (\frac{1}{2}\ln\frac{2}{\delta_m})^{-\frac{1}{2}} \delta_m^{-1}, \\
\frac{\partial^2 g}{\partial \delta_m^2} &= \frac{1}{4}\frac{1}{\alpha} n_m^{-\frac{1}{2}} (\frac{1}{2}\ln\frac{2}{\delta_m})^{-\frac{1}{2}} \delta_m^{-2} - \frac{1}{16}\frac{1}{\alpha} n_m^{-\frac{1}{2}} (\frac{1}{2}\ln\frac{2}{\delta_m})^{-\frac{3}{2}} \delta_m^{-2}.
\end{aligned}$$

We see that $\frac{\partial^2 g}{\partial n_m^2} > 0$.

What's more, we see that

$$\begin{aligned}
&|\nabla^2 g(n_m, \delta_m)| \\
&= \frac{\partial^2 g}{\partial n_m^2}\frac{\partial^2 g}{\partial \delta_m^2} - (\frac{\partial^2 g}{\partial n_m \partial \delta_m})^2 \\
&= \frac{3}{16}(\frac{1}{\alpha})^2 n_m^{-3}\delta_m^{-2} - \frac{3}{64}(\frac{1}{\alpha})^2 n_m^{-3}(\frac{1}{2}\ln\frac{2}{\delta_m})^{-1}\delta_m^{-2} - \frac{1}{64}(\frac{1}{\alpha})^2 n_m^{-3}(\frac{1}{2}\ln\frac{2}{\delta_m})^{-1}\delta_m^{-2} \\
&= \frac{3}{16}(\frac{1}{\alpha})^2 n_m^{-3}\delta_m^{-2} - \frac{1}{16}(\frac{1}{\alpha})^2 n_m^{-3}(\frac{1}{2}\ln\frac{2}{\delta_m})^{-1}\delta_m^{-2} \\
&= \frac{1}{16}(\frac{1}{\alpha})^2 n_m^{-3}\delta_m^{-2}\left(3 - (\frac{1}{2}\ln\frac{2}{\delta_m})^{-1}\right) \\
&> 0, \quad \forall \delta_m \in (0, \delta).
\end{aligned}$$

(Judging from the last formula above alone, " $>$ " holds as long as $\delta \in (0, 2/\exp(2/3)) = (0, 1.027)$. But $\delta < 1$ so it holds for all $\delta_m \in (0, \delta)$.)

Both leading principle minors are positive, so we have that the Hessian matrix of $g(n_m, \delta_m)$ is positive definite. Thus we have that $g(n_m, \delta_m)$ is convex.

Next look at the components of the Hessian matrix of $h(n_m, \delta_m)$:

$$\begin{aligned}
\frac{\partial^2 h}{\partial n_m^2} &= \frac{3}{4}\frac{1-\alpha}{\alpha}(n - n_m)^{-\frac{5}{2}} (\frac{1}{2}\ln\frac{2}{\delta - \delta_m})^{\frac{1}{2}}, \\
\frac{\partial^2 h}{\partial n_m \partial \delta_m} &= \frac{1}{8}\frac{1-\alpha}{\alpha}(n - n_m)^{-\frac{3}{2}} (\frac{1}{2}\ln\frac{2}{\delta - \delta_m})^{-\frac{1}{2}} (\delta - \delta_m)^{-1}, \\
\frac{\partial^2 h}{\partial \delta_m^2} &= \frac{1}{4}\frac{1-\alpha}{\alpha}(n - n_m)^{-\frac{1}{2}} (\frac{1}{2}\ln\frac{2}{\delta - \delta_m})^{-\frac{1}{2}} (\delta - \delta_m)^{-2} \\
&\quad -\frac{1}{16}\frac{1-\alpha}{\alpha}(n - n_m)^{-\frac{1}{2}} (\frac{1}{2}\ln\frac{2}{\delta - \delta_m})^{-\frac{3}{2}} (\delta - \delta_m)^{-2}.
\end{aligned}$$

We see that $\frac{\partial^2 h}{\partial n_m^2} > 0$.

Similar to above, we also observe that

$$
\begin{aligned}
&|\nabla^2 h(n_m, \delta_m)| \\
&= \frac{\partial^2 h}{\partial n_m^2}\frac{\partial^2 h}{\partial \delta_m^2} - (\frac{\partial^2 h}{\partial n_m \partial \delta_m})^2 \\
&= \frac{3}{16}(\frac{1-\alpha}{\alpha})^2 (n-n_m)^{-3}(\delta-\delta_m)^{-2} - \frac{3}{64}(\frac{1-\alpha}{\alpha})^2 (n-n_m)^{-3}(\frac{1}{2}\ln\frac{2}{\delta-\delta_m})^{-1}(\delta-\delta_m)^{-2} \\
&\quad - \frac{1}{64}(\frac{1-\alpha}{\alpha})^2 (n-n_m)^{-3}(\frac{1}{2}\ln\frac{2}{\delta-\delta_m})^{-1}(\delta-\delta_m)^{-2} \\
&= \frac{3}{16}(\frac{1-\alpha}{\alpha})^2 (n-n_m)^{-3}(\delta-\delta_m)^{-2} - \frac{1}{16}(\frac{1-\alpha}{\alpha})^2 (n-n_m)^{-3}(\frac{1}{2}\ln\frac{2}{\delta-\delta_m})^{-1}(\delta-\delta_m)^{-2} \\
&= \frac{1}{16}(\frac{1-\alpha}{\alpha})^2 (n-n_m)^{-3}(\delta-\delta_m)^{-2}\left(3 - (\frac{1}{2}\ln\frac{2}{\delta-\delta_m})^{-1}\right) \\
&> 0, \quad \forall \delta_m \in (0, \delta).
\end{aligned}
$$

(Judging from the last formula above alone, " $>$ " holds as long as $\delta \in (\delta - 2/\exp(2/3), \delta) = (\delta - 1.027, \delta)$. But $\delta < 1$ so it holds for all $\delta_m \in (0, \delta)$.)

Both leading principle minors are positive, so we have that the Hessian matrix of $h(n_m, \delta_m)$ is positive definite. Thus we have that $h(n_m, \delta_m)$ is convex.

Based on these two results, we have that $\epsilon(n_m, \delta_m) = g(n_m, \delta_m) + h(n_m, \delta_m)$ is convex in (n_m, δ_m).

Further, given that $\epsilon(n_m, \delta_m)$ is convex in (n_m, δ_m), the possible range of n_m is $(0, n)$ which is a convex nonempty set, we have

$$\epsilon(\delta_m) = \min_{n_m \in (0,n)} \epsilon(n_m, \delta_m)$$

also being convex in δ_m, since $\epsilon(\delta_m) > -\infty, \forall \delta_m \in (0, \delta)$. ■

Appendix B. Details of Isolation Forest Anomaly Score Calculation When Clean Data Set Contains Multiple Classes

When the clean data set contains multiple classes, on each class, an Isolation Forest computes 1000 full depth isolation trees, with each tree grown on a randomly-selected 20% subsample of the data points from this class. For each nominal point, we compute its anomaly score from the forest on its own class via out-of-bag estimates and its anomaly scores from other forests using the full forests respectively. For each point from the mixture data set, we compute its anomaly scores from all forests using the full forests respectively. The final anomaly score for each point is computed as the minimal of its anomaly scores from all forests.

Appendix C. Tables of Experimental Results in Section 7

C1. Synthetic Data Sets

In this section we include the simulation results on synthetic data sets from using Isolation Forest in Tables 3, 4, 5. For all cases, we include results from targeting on different recalls

Table 3: n^*, recall (i.e. alien detection rate) and false positive rate from experiments using 9-dimensional normal data, 98%, iForest

			Basic CDF			Iso CDF		
			Recall	False Positive Rate		Recall	False Positive Rate	
α	n	n^*	Recall±CI	FPR±CI	Oracle	Recall±CI	FPR±CI	Oracle
0.01	100	247818	0.710±0.033	0.033±0.027	0.102	0.929±0.029	0.512±0.080	0.102
	500	1167215	0.862±0.019	0.033±0.024	0.042	0.972±0.016	0.543±0.079	0.042
	1000	1829649	0.884±0.015	0.031±0.024	0.036	0.980±0.009	0.574±0.080	0.036
	5000	4236646	0.920±0.010	0.060±0.038	0.039	0.985±0.007	0.506±0.079	0.039
	10000	6363404	0.932±0.009	0.065±0.034	0.037	0.984±0.007	0.520±0.080	0.037
0.05	100	23373	0.826±0.027	0.088±0.037	0.101	0.950±0.022	0.502±0.081	0.101
	500	239656	0.939±0.009	0.064±0.032	0.042	0.979±0.007	0.465±0.081	0.042
	1000	259309	0.940±0.008	0.046±0.025	0.037	0.977±0.007	0.477±0.085	0.037
	5000	1067189	0.961±0.005	0.083±0.039	0.039	0.984±0.005	0.411±0.080	0.039
	10000	1536752	0.965±0.004	0.063±0.026	0.037	0.987±0.004	0.434±0.076	0.037
0.10	100	20178	0.907±0.017	0.105±0.033	0.100	0.977±0.010	0.549±0.075	0.100
	500	107381	0.951±0.007	0.071±0.035	0.042	0.985±0.005	0.482±0.080	0.042
	1000	196205	0.960±0.005	0.062±0.023	0.037	0.982±0.005	0.419±0.081	0.037
	5000	456821	0.970±0.004	0.075±0.031	0.039	0.988±0.004	0.403±0.075	0.039
	10000	861861	0.975±0.003	0.088±0.034	0.037	0.989±0.003	0.433±0.077	0.037
0.20	100	7550	0.946±0.011	0.158±0.045	0.101	0.974±0.010	0.496±0.075	0.101
	500	80449	0.971±0.005	0.131±0.045	0.042	0.988±0.004	0.484±0.078	0.042
	1000	110875	0.972±0.004	0.098±0.038	0.037	0.989±0.004	0.475±0.079	0.037
	5000	498016	0.977±0.002	0.048±0.010	0.039	0.985±0.003	0.254±0.066	0.039
	10000	670130	0.977±0.002	0.051±0.019	0.037	0.984±0.003	0.216±0.060	0.037
0.50	100	7053	0.970±0.005	0.156±0.036	0.102	0.982±0.005	0.395±0.073	0.102
	500	34712	0.977±0.003	0.056±0.009	0.042	0.984±0.003	0.256±0.065	0.042
	1000	70925	0.979±0.002	0.053±0.014	0.036	0.985±0.003	0.196±0.052	0.036
	5000	167019	0.978±0.001	0.039±0.002	0.039	0.979±0.001	0.049±0.014	0.039
	10000	451373	0.979±0.001	0.036±0.001	0.037	0.979±0.001	0.047±0.016	0.037

which are 98%, 95% and 90%. In Tables 3, 4, 5, the oracle FPR column is the mean of 100 oracle FPRs in each setting.

In Table 6, we include the results we used for plotting Figure 9. The results are the 1st quartile, median and 3rd quartile of FPR from experiments using iForest with target recall 95%. Here the oracle FPR column is the median of 100 oracle FPRs.

C2. Benchmark Data Sets

In this section we include results of performance on UCI benchmarks, MNIST and Tiny Imagenet and Tables 7-15 illustrate the results. The experimental protocol is similar to synthetic data sets. For Isolation forest we train Forest with 1000 trees on nominal data set and use out of bag estimates of this data set to estimate the nominal data sets anomaly score distribution. Tables 10-15 illustrate the results of iForest for 6 different data sets for

Table 4: n^*, recall (i.e. alien detection rate) and false positive rate from experiments using 9-dimensional normal data, 95%, iForest

			Basic CDF			Iso CDF		
			Recall	False Positive Rate		Recall	False Positive Rate	
α	n	n^*	Recall±CI	FPR±CI	Oracle	Recall±CI	FPR±CI	Oracle
0.01	100	275039	0.710±0.033	0.033±0.027	0.052	0.929±0.029	0.509±0.080	0.052
	500	1474209	0.862±0.019	0.033±0.024	0.015	0.972±0.016	0.533±0.079	0.015
	1000	2462157	0.884±0.015	0.030±0.024	0.012	0.978±0.010	0.557±0.081	0.012
	5000	6171393	0.911±0.010	0.039±0.030	0.014	0.982±0.008	0.496±0.080	0.014
	10000	9309633	0.918±0.010	0.054±0.032	0.014	0.981±0.008	0.495±0.081	0.014
0.05	100	27589	0.826±0.027	0.082±0.035	0.051	0.947±0.022	0.489±0.081	0.051
	500	243154	0.920±0.010	0.035±0.020	0.015	0.975±0.009	0.440±0.079	0.015
	1000	307512	0.923±0.009	0.022±0.011	0.012	0.966±0.010	0.420±0.084	0.012
	5000	1356124	0.943±0.005	0.040±0.028	0.014	0.973±0.007	0.351±0.079	0.014
	10000	1553411	0.945±0.005	0.024±0.009	0.014	0.972±0.006	0.314±0.074	0.014
0.10	100	28043	0.906±0.016	0.101±0.033	0.050	0.969±0.013	0.511±0.077	0.050
	500	109029	0.933±0.009	0.055±0.032	0.015	0.974±0.008	0.397±0.078	0.015
	1000	157112	0.934±0.006	0.017±0.006	0.012	0.969±0.007	0.313±0.075	0.012
	5000	1232102	0.949±0.004	0.027±0.018	0.014	0.967±0.006	0.194±0.061	0.014
	10000	861861	0.951±0.003	0.027±0.016	0.014	0.964±0.005	0.192±0.063	0.014
0.20	100	8666	0.929±0.012	0.126±0.042	0.051	0.963±0.013	0.428±0.073	0.051
	500	121266	0.953±0.006	0.054±0.025	0.015	0.977±0.006	0.360±0.075	0.015
	1000	177212	0.949±0.004	0.018±0.004	0.012	0.968±0.006	0.273±0.072	0.012
	5000	581132	0.949±0.002	0.014±0.001	0.014	0.953±0.003	0.039±0.024	0.014
	10000	776090	0.949±0.002	0.014±0.001	0.014	0.952±0.003	0.042±0.028	0.014
0.50	100	6349	0.952±0.006	0.084±0.021	0.052	0.966±0.007	0.262±0.061	0.052
	500	56529	0.951±0.003	0.018±0.002	0.015	0.954±0.004	0.038±0.021	0.015
	1000	111994	0.951±0.002	0.013±0.001	0.012	0.952±0.002	0.014±0.001	0.012
	5000	292413	0.950 ±0.001	0.014± 0.000	0.014	0.950±0.001	0.014±0.000	0.014
	10000	379279	0.950 ±0.001	0.014± 0.000	0.014	0.950± 0.001	0.014±0.000	0.014

Table 5: n^*, recall (i.e. alien detection rate) and false positive rate from experiments using 9-dimensional normal data, 90%, iForest

			Basic CDF			Iso CDF		
			Recall	False Positive Rate		Recall	False Positive Rate	
α	n	n^*	Recall±CI	FPR±CI	Oracle	Recall±CI	FPR±CI	Oracle
0.01	100	331513	0.710±0.033	0.033±0.027	0.026	0.929±0.029	0.509±0.080	0.026
	500	2340744	0.862±0.019	0.033±0.024	0.005	0.970±0.016	0.517±0.078	0.005
	1000	3222506	0.859±0.014	0.011±0.008	0.004	0.976±0.011	0.542±0.081	0.004
	5000	5918805	0.869±0.011	0.012±0.017	0.004	0.976±0.010	0.476±0.079	0.004
	10000	12543171	0.884±0.010	0.012±0.009	0.005	0.971±0.011	0.458±0.080	0.005
0.05	100	37658	0.826±0.027	0.081±0.034	0.026	0.936±0.024	0.468±0.081	0.026
	500	403920	0.893±0.011	0.020±0.015	0.006	0.960±0.012	0.372±0.075	0.006
	1000	482922	0.888±0.010	0.015±0.011	0.004	0.945±0.014	0.381±0.082	0.004
	5000	2307205	0.901±0.006	0.007±0.004	0.004	0.939±0.011	0.228±0.070	0.004
	10000	2629242	0.898±0.005	0.005±0.001	0.005	0.923±0.009	0.139±0.056	0.005
0.10	100	39085	0.879±0.017	0.059±0.021	0.025	0.957±0.016	0.463±0.076	0.025
	500	139647	0.900±0.010	0.019±0.014	0.005	0.944±0.013	0.297±0.073	0.005
	1000	156669	0.888±0.008	0.005±0.001	0.004	0.925±0.012	0.166±0.058	0.004
	5000	1867515	0.902±0.003	0.004±0.000	0.003	0.911±0.006	0.060±0.039	0.003
	10000	1232102	0.900±0.003	0.005±0.000	0.005	0.903±0.004	0.016±0.015	0.005
0.20	100	6481	0.881±0.017	0.060±0.022	0.026	0.942±0.016	0.359±0.072	0.026
	500	63235	0.909±0.007	0.010±0.003	0.005	0.937±0.010	0.170±0.057	0.005
	1000	153077	0.902±0.004	0.005±0.000	0.004	0.913±0.007	0.066±0.040	0.004
	5000	397467	0.898±0.002	0.003±0.000	0.004	0.898±0.002	0.004±0.000	0.004
	10000	1088542	0.899±0.002	0.005±0.000	0.005	0.900±0.002	0.005±0.000	0.005
0.50	100	4400	0.912±0.008	0.038±0.005	0.026	0.920±0.010	0.107±0.042	0.026
	500	22825	0.904±0.004	0.006±0.000	0.005	0.904±0.004	0.006±0.000	0.005
	1000	44373	0.903±0.003	0.004±0.000	0.004	0.903±0.003	0.004±0.000	0.004
	5000	229795	0.900±0.001	0.004±0.000	0.004	0.900±0.001	0.004±0.000	0.004
	10000	374065	0.900±0.001	0.005±0.000	0.005	0.900±0.001	0.005±0.000	0.005

Table 6: 1st quartile, median, 3rd quartile of false positive rate from experiments using 9-dimensional normal data, 95%, Iforest

		Basic CDF False Positive Rate			
α	n	1st quartile	median	3rd quartile	Oracle(median)
0.01	100	0.004	0.006	0.014	0.051
	500	0.002	0.004	0.015	0.015
	1000	0.002	0.004	0.010	0.012
	5000	0.002	0.004	0.013	0.014
	10000	0.003	0.006	0.014	0.014
0.05	100	0.006	0.014	0.043	0.050
	500	0.004	0.008	0.023	0.015
	1000	0.004	0.008	0.015	0.012
	5000	0.006	0.010	0.020	0.014
	10000	0.008	0.011	0.019	0.014
0.1	100	0.015	0.032	0.094	0.049
	500	0.006	0.012	0.021	0.015
	1000	0.005	0.009	0.014	0.012
	5000	0.009	0.013	0.020	0.014
	10000	0.011	0.014	0.017	0.014
0.2	100	0.025	0.043	0.105	0.049
	500	0.010	0.018	0.031	0.015
	1000	0.008	0.011	0.018	0.012
	5000	0.010	0.013	0.015	0.014
	10000	0.012	0.013	0.015	0.014
0.5	100	0.040	0.058	0.090	0.051
	500	0.012	0.016	0.021	0.015
	1000	0.011	0.012	0.016	0.012
	5000	0.013	0.014	0.016	0.014
	10000	0.013	0.014	0.015	0.014

Table 7: Recall (i.e. alien detection rate) & false positive rate for Image data sets,98%

			Basic CDF	
			Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	recall±CI	FPR±CI
Tiny Image Net	0.100	0.100	0.932 ± 0.012	0.690 ± 0.027
n=5 263	0.100	0.104	0.942 ± 0.011	0.715 ± 0.027
	0.100	0.108	0.952 ± 0.009	0.739 ± 0.027
	0.200	0.200	0.963 ± 0.006	0.744 ± 0.019
	0.200	0.204	0.967 ± 0.005	0.758 ± 0.019
	0.200	0.208	0.971 ± 0.005	0.772 ± 0.019
	0.400	0.400	0.976 ± 0.003	0.767 ± 0.011
	0.400	0.404	0.978 ± 0.003	0.775 ± 0.011
	0.400	0.408	0.979 ± 0.002	0.783 ± 0.010
mnist	0.100	0.100	0.956 ± 0.008	0.406 ± 0.034
n=10 000	0.100	0.104	0.968 ± 0.007	0.451 ± 0.034
	0.100	0.108	0.980 ± 0.005	0.511 ± 0.036
	0.200	0.200	0.971 ± 0.004	0.425 ± 0.025
	0.200	0.204	0.980 ± 0.004	0.471 ± 0.029
	0.200	0.208	0.986 ± 0.003	0.519 ± 0.031
	0.400	0.400	0.979 ± 0.002	0.407 ± 0.009
	0.400	0.404	0.984 ± 0.002	0.433 ± 0.012
	0.400	0.408	0.988 ± 0.002	0.462 ± 0.015

varying values of η and report the observed recall, false positive rate averaged over 100 runs of each experiment.

For Image data sets we follow the same protocol as UCI for MNIST and apply Isolation Forest on the input image but for Tiny Imagenet the anomaly scores are obtained differently. We first train a Wide Residual Network (40-2) classifier on the 200 nominal classes of Tiny Imagenet and apply baseline method Hendrycks and Gimpel (2017) on validation data to get the nominal data set distribution and later apply the same method on the mixture data set which will have α proportion of aliens which are basically from 800 held out classes. Tables 7-9 illustrate the results for these data sets for target recall of 98%, 95% and 90%.

Table 8: Recall (i.e. alien detection rate) & false positive rate for Image data sets,95%

			Basic CDF	
			Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	recall±CI	FPR±CI
Tiny Image Net	0.100	0.100	0.912 ± 0.013	0.638 ± 0.025
n=5 263	0.100	0.104	0.922 ± 0.012	0.658 ± 0.025
	0.100	0.108	0.932 ± 0.011	0.678 ± 0.025
	0.200	0.200	0.937 ± 0.007	0.665 ± 0.015
	0.200	0.204	0.943 ± 0.006	0.678 ± 0.015
	0.200	0.208	0.948 ± 0.006	0.691 ± 0.015
	0.400	0.400	0.949 ± 0.003	0.673 ± 0.006
	0.400	0.404	0.951 ± 0.003	0.679 ± 0.006
	0.400	0.408	0.954 ± 0.003	0.686 ± 0.007
MNIST	0.100	0.100	0.936 ± 0.009	0.340 ± 0.026
n=10 000	0.100	0.104	0.953 ± 0.008	0.385 ± 0.029
	0.100	0.108	0.966 ± 0.007	0.427 ± 0.031
	0.200	0.200	0.946 ± 0.005	0.328 ± 0.011
	0.200	0.204	0.957 ± 0.005	0.356 ± 0.014
	0.200	0.208	0.967 ± 0.004	0.388 ± 0.016
	0.400	0.400	0.950 ± 0.002	0.322 ± 0.004
	0.400	0.404	0.956 ± 0.002	0.335 ± 0.005
	0.400	0.408	0.962 ± 0.002	0.348 ± 0.005

Table 9: Recall (i.e. alien detection rate) & false positive rate for Image data sets,90%

			Basic CDF	
			Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	recall±CI	FPR±CI
Tiny Image Net	0.100	0.100	0.874 ± 0.014	0.564 ± 0.021
n=5 263	0.100	0.104	0.884 ± 0.013	0.580 ± 0.020
	0.100	0.108	0.895 ± 0.013	0.596 ± 0.021
	0.200	0.200	0.894 ± 0.007	0.577 ± 0.010
	0.200	0.204	0.900 ± 0.007	0.586 ± 0.010
	0.200	0.208	0.905 ± 0.007	0.596 ± 0.010
	0.400	0.400	0.898 ± 0.003	0.578 ± 0.004
	0.400	0.404	0.902 ± 0.003	0.583 ± 0.004
	0.400	0.408	0.905 ± 0.003	0.587 ± 0.004
MNIST	0.100	0.100	0.891 ± 0.010	0.253 ± 0.014
n=10 000	0.100	0.104	0.916 ± 0.009	0.287 ± 0.017
	0.100	0.108	0.935 ± 0.008	0.327 ± 0.021
	0.200	0.200	0.897 ± 0.004	0.246 ± 0.005
	0.200	0.204	0.910 ± 0.005	0.262 ± 0.005
	0.200	0.208	0.921 ± 0.005	0.279 ± 0.006
	0.400	0.400	0.899 ± 0.002	0.246 ± 0.002
	0.400	0.404	0.906 ± 0.002	0.253 ± 0.002
	0.400	0.408	0.912 ± 0.002	0.261 ± 0.002

Table 10: Recall & false positive rate for Letter Recognition data set using Iforest for varying q (target recall $1 - q$)

				Basic CDF	
				Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	q	recall±CI	FPR±CI
Letter recognition	0.100	0.100	0.020	0.914 ± 0.014	0.277 ± 0.035
n=802	0.100	0.104	0.020	0.927 ± 0.013	0.314 ± 0.038
	0.100	0.108	0.020	0.940 ± 0.011	0.352 ± 0.042
	0.200	0.200	0.020	0.944 ± 0.009	0.316 ± 0.034
	0.200	0.204	0.020	0.951 ± 0.008	0.343 ± 0.036
	0.200	0.208	0.020	0.959 ± 0.008	0.373 ± 0.039
	0.400	0.400	0.020	0.965 ± 0.005	0.336 ± 0.025
	0.400	0.404	0.020	0.970 ± 0.004	0.364 ± 0.027
	0.400	0.408	0.020	0.974 ± 0.004	0.393 ± 0.029
	0.100	0.100	0.050	0.898 ± 0.015	0.236 ± 0.030
	0.100	0.104	0.050	0.916 ± 0.013	0.274 ± 0.034
	0.100	0.108	0.050	0.928 ± 0.012	0.305 ± 0.036
	0.200	0.200	0.050	0.928 ± 0.009	0.259 ± 0.027
	0.200	0.204	0.050	0.937 ± 0.009	0.285 ± 0.029
	0.200	0.208	0.050	0.944 ± 0.008	0.303 ± 0.031
	0.400	0.400	0.050	0.943 ± 0.005	0.244 ± 0.013
	0.400	0.404	0.050	0.949 ± 0.005	0.260 ± 0.014
	0.400	0.408	0.050	0.954 ± 0.005	0.279 ± 0.018
	0.100	0.100	0.100	0.870 ± 0.017	0.192 ± 0.024
	0.100	0.104	0.100	0.889 ± 0.016	0.215 ± 0.027
	0.100	0.108	0.100	0.902 ± 0.014	0.242 ± 0.029
	0.200	0.200	0.100	0.894 ± 0.011	0.187 ± 0.014
	0.200	0.204	0.100	0.904 ± 0.010	0.203 ± 0.016
	0.200	0.208	0.100	0.914 ± 0.009	0.220 ± 0.019
	0.400	0.400	0.100	0.902 ± 0.006	0.170 ± 0.004
	0.400	0.404	0.100	0.908 ± 0.005	0.178 ± 0.004
	0.400	0.408	0.100	0.914 ± 0.005	0.184 ± 0.005

Table 11: Recall & false positive rate for page.blocks data set using iForest for varying q (target recall $1 - q$)

				Basic CDF	
				Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	q	recall±CI	FPR±CI
pageblocks	0.100	0.100	0.020	0.921 ± 0.014	0.223 ± 0.036
n=1 112	0.100	0.104	0.020	0.937 ± 0.012	0.256 ± 0.039
	0.100	0.108	0.020	0.953 ± 0.011	0.291 ± 0.041
	0.200	0.200	0.020	0.954 ± 0.008	0.239 ± 0.030
	0.200	0.204	0.020	0.964 ± 0.007	0.272 ± 0.035
	0.200	0.208	0.020	0.971 ± 0.006	0.307 ± 0.039
	0.400	0.400	0.020	0.971 ± 0.004	0.229 ± 0.018
	0.400	0.404	0.020	0.977 ± 0.003	0.259 ± 0.025
	0.400	0.408	0.020	0.982 ± 0.003	0.289 ± 0.028
	0.100	0.100	0.050	0.905 ± 0.015	0.188 ± 0.029
	0.100	0.104	0.050	0.924 ± 0.014	0.219 ± 0.033
	0.100	0.108	0.050	0.939 ± 0.012	0.254 ± 0.037
	0.200	0.200	0.050	0.935 ± 0.009	0.186 ± 0.022
	0.200	0.204	0.050	0.946 ± 0.008	0.207 ± 0.023
	0.200	0.208	0.050	0.956 ± 0.008	0.233 ± 0.028
	0.400	0.400	0.050	0.947 ± 0.005	0.160 ± 0.009
	0.400	0.404	0.050	0.955 ± 0.004	0.172 ± 0.010
	0.400	0.408	0.050	0.961 ± 0.004	0.185 ± 0.011
	0.100	0.100	0.100	0.867 ± 0.016	0.142 ± 0.020
	0.100	0.104	0.100	0.891 ± 0.016	0.164 ± 0.023
	0.100	0.108	0.100	0.908 ± 0.015	0.192 ± 0.026
	0.200	0.200	0.100	0.893 ± 0.010	0.131 ± 0.010
	0.200	0.204	0.100	0.907 ± 0.009	0.141 ± 0.011
	0.200	0.208	0.100	0.920 ± 0.009	0.157 ± 0.015
	0.400	0.400	0.100	0.899 ± 0.004	0.121 ± 0.005
	0.400	0.404	0.100	0.907 ± 0.004	0.125 ± 0.006
	0.400	0.408	0.100	0.916 ± 0.004	0.129 ± 0.006

Table 12: Recall & false positive rate for Optical.digits data set using Iforest for varying q (target recall $1 - q$)

				Basic CDF	
				Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	q	recall±CI	FPR±CI
Optical.digits	0.100	0.100	0.020	0.926 ± 0.012	0.268 ± 0.040
n=1 492	0.100	0.104	0.020	0.940 ± 0.011	0.306 ± 0.043
	0.100	0.108	0.020	0.954 ± 0.009	0.349 ± 0.045
	0.200	0.200	0.020	0.953 ± 0.007	0.291 ± 0.034
	0.200	0.204	0.020	0.963 ± 0.006	0.328 ± 0.037
	0.200	0.208	0.020	0.970 ± 0.006	0.366 ± 0.040
	0.400	0.400	0.020	0.972 ± 0.004	0.313 ± 0.028
	0.400	0.404	0.020	0.977 ± 0.003	0.345 ± 0.031
	0.400	0.408	0.020	0.982 ± 0.003	0.380 ± 0.034
	0.100	0.100	0.050	0.908 ± 0.013	0.231 ± 0.036
	0.100	0.104	0.050	0.928 ± 0.012	0.265 ± 0.039
	0.100	0.108	0.050	0.942 ± 0.011	0.296 ± 0.039
	0.200	0.200	0.050	0.934 ± 0.009	0.226 ± 0.026
	0.200	0.204	0.050	0.943 ± 0.008	0.250 ± 0.028
	0.200	0.208	0.050	0.953 ± 0.007	0.280 ± 0.031
	0.400	0.400	0.050	0.949 ± 0.004	0.216 ± 0.015
	0.400	0.404	0.050	0.955 ± 0.004	0.229 ± 0.016
	0.400	0.408	0.050	0.961 ± 0.004	0.248 ± 0.018
	0.100	0.100	0.100	0.874 ± 0.015	0.179 ± 0.028
	0.100	0.104	0.100	0.897 ± 0.014	0.205 ± 0.031
	0.100	0.108	0.100	0.914 ± 0.013	0.236 ± 0.034
	0.200	0.200	0.100	0.893 ± 0.009	0.159 ± 0.015
	0.200	0.204	0.100	0.906 ± 0.009	0.173 ± 0.017
	0.200	0.208	0.100	0.918 ± 0.009	0.190 ± 0.019
	0.400	0.400	0.100	0.900 ± 0.004	0.145 ± 0.003
	0.400	0.404	0.100	0.907 ± 0.004	0.152 ± 0.003
	0.400	0.408	0.100	0.915 ± 0.004	0.159 ± 0.004

Table 13: Recall & false positive rate for Landsat data set using Iforest for varying q (target recall $1-q$)

				Basic CDF	
				Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	q	recall±CI	FPR±CI
Landsat	0.100	0.100	0.020	0.917 ± 0.012	0.401 ± 0.048
n=1 600	0.100	0.104	0.020	0.927 ± 0.011	0.443 ± 0.047
	0.100	0.108	0.020	0.938 ± 0.010	0.489 ± 0.045
	0.200	0.200	0.020	0.943 ± 0.008	0.489 ± 0.037
	0.200	0.204	0.020	0.950 ± 0.007	0.522 ± 0.036
	0.200	0.208	0.020	0.957 ± 0.007	0.552 ± 0.035
	0.400	0.400	0.020	0.967 ± 0.005	0.568 ± 0.025
	0.400	0.404	0.020	0.971 ± 0.004	0.586 ± 0.025
	0.400	0.408	0.020	0.974 ± 0.004	0.602 ± 0.024
	0.100	0.100	0.050	0.902 ± 0.012	0.348 ± 0.045
	0.100	0.104	0.050	0.916 ± 0.012	0.389 ± 0.046
	0.100	0.108	0.050	0.926 ± 0.011	0.426 ± 0.044
	0.200	0.200	0.050	0.926 ± 0.008	0.407 ± 0.035
	0.200	0.204	0.050	0.934 ± 0.008	0.438 ± 0.034
	0.200	0.208	0.050	0.940 ± 0.008	0.464 ± 0.034
	0.400	0.400	0.050	0.942 ± 0.005	0.447 ± 0.022
	0.400	0.404	0.050	0.946 ± 0.005	0.465 ± 0.021
	0.400	0.408	0.050	0.951 ± 0.005	0.484 ± 0.021
	0.100	0.100	0.100	0.872 ± 0.013	0.258 ± 0.039
	0.100	0.104	0.100	0.892 ± 0.012	0.303 ± 0.040
	0.100	0.108	0.100	0.904 ± 0.011	0.341 ± 0.041
	0.200	0.200	0.100	0.893 ± 0.009	0.283 ± 0.030
	0.200	0.204	0.100	0.903 ± 0.008	0.311 ± 0.030
	0.200	0.208	0.100	0.910 ± 0.008	0.336 ± 0.030
	0.400	0.400	0.100	0.900 ± 0.005	0.286 ± 0.018
	0.400	0.404	0.100	0.905 ± 0.005	0.303 ± 0.018
	0.400	0.408	0.100	0.910 ± 0.005	0.321 ± 0.018

Table 14: Recall & False Positive Rate for Shuttle Data set using iForest for varying q (target recall $1 - q$)

				Basic CDF	
				Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	q	recall±CI	FPR±CI
Shuttle	0.100	0.100	0.020	0.979 ± 0.001	0.022 ± 0.016
n=8 777	0.100	0.104	0.020	0.997 ± 0.001	0.179 ± 0.044
	0.100	0.108	0.020	1.000 ± 0.000	0.339 ± 0.053
	0.200	0.200	0.020	0.980 ± 0.001	0.010 ± 0.003
	0.200	0.204	0.020	0.995 ± 0.001	0.087 ± 0.026
	0.200	0.208	0.020	1.000 ± 0.000	0.244 ± 0.044
	0.400	0.400	0.020	0.980 ± 0.000	0.005 ± 0.000
	0.400	0.404	0.020	0.990 ± 0.001	0.025 ± 0.010
	0.400	0.408	0.020	0.997 ± 0.000	0.080 ± 0.019
	0.100	0.100	0.050	0.952 ± 0.001	0.002 ± 0.001
	0.100	0.104	0.050	0.986 ± 0.001	0.036 ± 0.016
	0.100	0.108	0.050	0.998 ± 0.001	0.174 ± 0.039
	0.200	0.200	0.050	0.951 ± 0.000	0.001 ± 0.000
	0.200	0.204	0.050	0.969 ± 0.001	0.002 ± 0.000
	0.200	0.208	0.050	0.987 ± 0.001	0.024 ± 0.008
	0.400	0.400	0.050	0.951 ± 0.000	0.001 ± 0.000
	0.400	0.404	0.050	0.960 ± 0.000	0.001 ± 0.000
	0.400	0.408	0.050	0.969 ± 0.000	0.001 ± 0.000
	0.100	0.100	0.100	0.903 ± 0.001	0.001 ± 0.000
	0.100	0.104	0.100	0.938 ± 0.001	0.001 ± 0.000
	0.100	0.108	0.100	0.972 ± 0.001	0.008 ± 0.005
	0.200	0.200	0.100	0.902 ± 0.001	0.001 ± 0.000
	0.200	0.204	0.100	0.919 ± 0.000	0.001 ± 0.000
	0.200	0.208	0.100	0.937 ± 0.000	0.001 ± 0.000
	0.400	0.400	0.100	0.901 ± 0.000	0.001 ± 0.000
	0.400	0.404	0.100	0.910 ± 0.000	0.001 ± 0.000
	0.400	0.408	0.100	0.919 ± 0.000	0.001 ± 0.000

Table 15: Recall & False Positive Rate for Covertype data set using Iforest for varying q (target recall $1-q$)

				Basic CDF	
				Recall	False Positive Rate
Data set	α	$\hat{\alpha}$	q	recall±CI	FPR±CI
Covertype	0.100	0.100	0.020	0.945 ± 0.009	0.851 ± 0.020
n=10 000	0.100	0.104	0.020	0.951 ± 0.008	0.865 ± 0.018
	0.100	0.108	0.020	0.957 ± 0.007	0.879 ± 0.016
	0.200	0.200	0.020	0.970 ± 0.004	0.908 ± 0.009
	0.200	0.204	0.020	0.972 ± 0.003	0.913 ± 0.008
	0.200	0.208	0.020	0.974 ± 0.003	0.918 ± 0.007
	0.400	0.400	0.020	0.978 ± 0.001	0.926 ± 0.003
	0.400	0.404	0.020	0.978 ± 0.001	0.928 ± 0.003
	0.400	0.408	0.020	0.979 ± 0.001	0.929 ± 0.003
	0.100	0.100	0.050	0.919 ± 0.010	0.790 ± 0.021
	0.100	0.104	0.050	0.925 ± 0.010	0.803 ± 0.020
	0.100	0.108	0.050	0.932 ± 0.009	0.816 ± 0.018
	0.200	0.200	0.050	0.940 ± 0.004	0.832 ± 0.010
	0.200	0.204	0.050	0.942 ± 0.004	0.838 ± 0.010
	0.200	0.208	0.050	0.944 ± 0.004	0.844 ± 0.009
	0.400	0.400	0.050	0.948 ± 0.002	0.850 ± 0.004
	0.400	0.404	0.050	0.949 ± 0.002	0.853 ± 0.004
	0.400	0.408	0.050	0.950 ± 0.002	0.855 ± 0.004
	0.100	0.100	0.100	0.879 ± 0.011	0.708 ± 0.020
	0.100	0.104	0.100	0.887 ± 0.011	0.721 ± 0.019
	0.100	0.108	0.100	0.893 ± 0.010	0.732 ± 0.018
	0.200	0.200	0.100	0.894 ± 0.006	0.729 ± 0.012
	0.200	0.204	0.100	0.897 ± 0.005	0.735 ± 0.011
	0.200	0.208	0.100	0.900 ± 0.005	0.742 ± 0.010
	0.400	0.400	0.100	0.899 ± 0.002	0.740 ± 0.005
	0.400	0.404	0.100	0.900 ± 0.002	0.743 ± 0.005
	0.400	0.408	0.100	0.902 ± 0.002	0.746 ± 0.005

References

R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

A. Bendale and T. E. Boult. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572, June 2016.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010.

H. Cevikalp and B. Triggs. Efficient object detection using cascades of nearest convex model classifiers. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3138–3145, June 2012.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, Jan 1970. ISSN 0018-9448.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016.

Qing Da, Yang Yu, and Zhi-Hua Zhou. Learning with augmented class by exploiting unlabeled data. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1760–1766. AAAI Press, 2014.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems*, pages 9157–9168, 2018.

Andrew F Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, pages 16–21. ACM, 2013.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4885–4894. Curran Associates, Inc., 2017.

B. Heflin, W. Scheirer, and T. E. Boult. Detecting and classifying scars, marks, and tattoos found in the wild. In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 31–38, Sept 2012.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Hongliang Jin, Qingshan Liu, and Hanqing Lu. Face detection using one-class-based support vectors. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 457–462, May 2004.

Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V Pazinato, Waldir R de Almeida, Otávio AB Penatti, Ricardo da S Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, 2017.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018.

Zhenfeng Lin and James P Long. A flexible procedure for mixture proportion estimation in positive-unlabeled learning. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(2):178–187, 2020.

Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. Partially supervised classification of text documents. In *ICML*, volume 2, pages 387–394, 2002.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422. IEEE, 2008.

Si Liu, Risheek Garrepalli, Thomas Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with PAC guarantees. In *International Conference on Machine Learning*, pages 3169–3178, 2018.

David A Lytle, Gonzalo Martínez-Muñoz, Wei Zhang, Natalia Larios, Linda Shapiro, Robert Paasch, Andrew Moldenke, Eric N Mortensen, Sinisa Todorovic, and Thomas G Dietterich. Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society*, 29(3):867–874, 2010.

Larry M. Manevitz and Malik Yousef. One-class SVMs for document classification. *J. Mach. Learn. Res.*, 2:139–154, March 2002. ISSN 1532-4435.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. ISSN 00911798.

Aditya Krishna Menon and Robert C Williamson. A loss framework for calibrated anomaly detection. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1494–1504. Curran Associates Inc., 2018.

Rohit Kumar Patra and Bodhisattva Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2016.

Tadeusz Pietraszek. Optimizing abstaining classifiers using ROC analysis. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 665–672, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5.

Dimitrios A. Pritsos and Efstathios Stamatatos. *Open-Set Classification for Automated Genre Identification*, pages 207–217. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-36973-5.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pages 2052–2060, 2016.

Tyler Sanderson and Clayton Scott. Class proportion estimation with application to multiclass anomaly rejection. In *Artificial Intelligence and Statistics*, pages 850–858, 2014.

W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, July 2013. ISSN 0162-8828.

W. J. Scheirer, L. P. Jain, and T. E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, Nov 2014. ISSN 0162-8828.

Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, July 2001. ISSN 0899-7667.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pages 838–846, 2015.

Lei Shu, Hu Xu, and Bing Liu. DOC: deep open classification of text documents. *CoRR*, abs/1709.08716, 2017.

D.M.J. Tax and R.P.W. Duin. Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29(10):1565 – 1570, 2008. ISSN 0167-8655.

Marten H. Wegkamp. Lasso type classifiers with a reject option. 2007.

M. Wu and J. Ye. A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2088–2092, Nov 2009. ISSN 0162-8828.

Xiang Sean Zhou and Thomas S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, Apr 2003. ISSN 1432-1882.