The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Can We Achieve Open Category Detection with Guarantees?

Si Liu*
Department of Statistics
Oregon State University
lius2@oregonstate.edu

Risheek Garrepalli*
School of EECS
Oregon State University
garrepar@oregonstate.edu

Alan Fern
School of EECS
Oregon State University
Alan.Fern@oregonstate.edu

Thomas G. Dietterich
School of EECS
Oregon State University
tgd@oregonstate.edu

Abstract

Open category detection is the problem of detecting "alien" test instances that belong to categories/classes that were not present in the training data. In many applications, reliably detecting such aliens is central to ensuring safety and/or quality of test data analysis. Unfortunately, to the best of our knowledge, 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 simplified, 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. However, at test time, some fraction α of the test examples are aliens. Under the assumption that we know an upper bound on α, we develop an algorithm with PAC-style guarantees on the alien detection rate, while aiming to minimize false alarms. Our empirical results on synthetic and benchmark datasets demonstrate the regimes in which the algorithm can be effective and provide a baseline for further advancements.

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" data, can arise without warning. Ignoring the potential for such aliens can lead to serious safety concerns in many applications and significantly degrade the quality of test data analysis in others. For example, consider a scientific application where a classifier is learned to recognize specific categories of insects in freshwater samples in order to detect important environmental changes (Lytle et al. 2010). Test samples will typically contain some number of species that are not represented in the training data. A classifier that is unaware of these new species will count them as existing species, leading to incorrect statistics and unreliable scientific conclusions.

The problem of open category detection is to detect such alien categories at test time. An ideal algorithm for this problem would guarantee a user-specified alien-detection rate (e.g. 95%), while attempting to minimize the false-alarm rate. Unfortunately, to the best of our knowledge, there are no algorithms that provide such guarantees under general conditions. In addition, empirical evaluations of existing algorithms for open category detection typically do not directly evaluate alien detection rates, which are perhaps the most relevant for safety-critical applications. Overall, our current theoretical and practical understanding of open category detection is lacking from a safety and quality perspective.

Is it possible to achieve open category detection with guarantees? In this paper, we take a step toward answering this question by studying a simplified, but practically relevant, problem setting. To motivate our setting, consider the above insect identification problem. At training time it is reasonable to expect that a clean training set is available that contains only the insect categories of interest. At test time, a new sample will include insects from the training categories along with some percentage of insects from new alien categories. Further, scientists may have reasonable estimates for this percentage based on their scientific knowledge and practical experience. We would like to guarantee that the system is able to raise an alarm for 95% of the insects from alien classes, with each alarm being examined by a scientist. At the same time, we would like to avoid as many "false alarms" as possible, since each alarm requires scientist effort.

To formalize the example, our setting assumes a clean training data set involving a finite set of categories and a test data set that contains a fraction α of aliens. Our first contribution is show that, in this setting, theoretical guarantees are possible given knowledge of an upper bound on α.

*These authors contributed equally to the work.

In particular, we give an algorithm that uses this knowledge in order 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 and prior data.

The key idea behind our algorithm is to leverage modern anomaly detectors, which are trained on the clean data. Our algorithm combines the anomaly-score distributions over the training and testing data in order to derive an alarm threshold rule that achieves the desired guarantee on the alien detection rate. In theory the detection rate guarantee will be met regardless of the quality of the anomaly detector. The quality of the detector, however, has a significant impact on the false alarm rate, with better detectors leading to fewer false alarms. Here we define "false alarm rate" as the fraction of alarms that are false, and it is different from, but closely related to the more common definition which is the fraction of nominal data points incorrect classified as aliens.

We carry out experiments on synthetic and benchmark datasets using a state-of-the-art anomaly detector, the Isolation Forest (Liu, Ting, and Zhou 2008). We vary the amount of training data, the fraction α of alien data points, along with the accuracy of the upper bound on α provided to our algorithm. 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 datasets and/or small values of α. Overall, our results provide a useful baseline for driving future work on open category detection with guarantees.

Related Work

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) (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 (Zhou and Huang 2003), which has been partly addressed by later work (Manevitz and Yousef 2002; Wu and Ye 2009; Jin, Liu, and Lu 2004; Cevikalp and Triggs 2012). OCSVMs have been employed in a multi-class setting similar to open category detection (Heflin, Scheirer, and Boult 2012; Pritsos and Stamatatos 2013). However, there are no direct mechanisms to control the alien detection rate of these methods, which is a key requirement for our problem setting.

Work on classification with rejection/abstaining options (Chow 1970; Wegkamp 2007; Tax and Duin 2008; Pietraszek 2005) 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 on formalizing the concept of "open space" to characterize the region of the feature space outside of the support of the training set (Scheirer et al. 2013). Variants of SVMs have also been developed, such as the One-vs-Set Machine (Scheirer et al. 2013) and the Weibull-calibrated SVM (Scheirer, Jain, and Boult 2014).Additional work has addressed open category detection by tuning the decision boundary based on unlabeled data which contains data from novel categories (Da, Yu, and Zhou 2014). Approaches based on nearest neighbor methods have also been proposed (Mendes Júnior et al. 2017). None of these methods, however, allow for the direct control of alien detection rates nor do they provide theoretical guarantees.

There is also a recent interest in open category detection for deep neural networks applied to vision and text classification (Bendale and Boult 2016; Shu, Xu, and Liu 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 infer aliens. 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 (Hendrycks and Gimpel 2016; Liang, Li, and Srikant 2017). All of this work is quite specialized to deep neural networks and still does not provide direct control of alien detection rates or theoretical guarantees.

Problem Setup

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 based on k i.i.d. draws from D_0. In practice, S_0 will correspond to some curated labeled data that contains only known categories of interest. We also receive as input an unlabeled "mixture" test 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 anomaly distribution D_a, which is a distribution over novel categories, or alien data points. We assume that D_a is stationary, so that all anomalies that appear as future test queries will also be drawn from D_a.

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 alien test points are drawn from D_a.

Given these data sets, our problem is to label test instances in D_m as either "alien" or "nominal". In particular, we wish to achieve a specified alien detection rate, that is the fraction of alien data in D_m that are classified as "alien" (e.g., 95%).

At the same time we would like the false alarm rate to be small, which is the fraction of alarms that are false.

Our approach to this problem will assume the availability of an anomaly detector, which is trained on S_0 and assigns anomaly scores to all data points in both S_0 and S_m. Intuitively, the anomaly scores order the data 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 on the anomaly scores and labeling all data with scores above the threshold as aliens and others as nominals. Our key challenge is to select a threshold that provides a guarantee on the alien detection rate.

Algorithms for Open Category Detection

In order to obtain theoretical guarantees, our algorithm assumes knowledge of the alien mixture probability α that is used to generate the mixture data S_m. Later, we will show that simply knowing an upper bound on α is sufficient for guarantees.

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 derive F_a. Rather, we have samples S_m and S_0 to work with. Thus, our algorithms work with empirical CDFs $\hat{F}_0$ and $\hat{F}_m$ which are simple step-wise approximations, and estimate an empirical CDF over aliens:

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

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. The steps of this algorithm are as follows.

Although $\hat{F}_m$ and $\hat{F}_0$ are both legal CDFs, the estimate for $\hat{F}_a$ from step 3 in the algorithm above may not necessarily be a legal CDF—it may not increase monotonically and it may even be negative. One common technique for dealing with this problem is to use isotonization (Barlow and Brunk 1972) and clipping. Isotonization finds the monotonically increasing function $\hat{F}_a^*$ that is closest to $\hat{F}_a$ in squared error. To convert $\hat{F}_a$ into a legal CDF, we first isotonize $\hat{F}_a$, and then for the isotonized version, set the part smaller than 0 to 0 and the part greater than 1 to 1. Denote this this isotonized and clipped version of the CDF as $\check{F}_a$. In our experiments, we consider a variant of the above algorithm that uses $\check{F}_a$ instead of $\hat{F}_a$ in step 4.

Algorithm 1

1: Get anomaly scores for all points in S_0 and S_m denoted $x_1, x_2, \cdots, x_k$ and $y_1, y_2, \cdots, y_n$ 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 $S = \{x_1, x_2, \cdots, x_k, y_1, y_2, \cdots, y_n\}$.

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 and derive bounds on amount of data required to achieve performance guarantees. Since our algorithm depends on random i.i.d. samples of nominal and mixture data our guarantees are probabilistic in nature. In particular, our goal is to derive a sufficient value for n that will guarantee with high probability over runs of the algorithm that a specified alien alarm rate of $1-q$ is approximately achieved.

For simplicity we will assume that the sizes of the nominal and mixture datasets are equal, that is, $|S_0| = |S_m| = n$. The key theoretical tool we use is a finite sample result on the uniform convergence of empirical CDF functions (Massart 1990). To use this result, we will make a reasonable technical assumption that the nominal and alien CDFs F_0 and F_a are continuous with convex support. In the following, let $1-q$ be the target alien detection rate and τ_q be the true q-quantile of the alien CDF. Any threshold lower than τ_q will guarantee an alien detection rate of at least $1-q$.

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 $1-q-\epsilon$.

The proof is in the Appendix. Qualitatively we see that the required sample size n grows logarithmically in $1/\delta$, which is not a major concern. More significantly, n must grow as

$\left(\frac{1}{\alpha}\right)^2$ and $\left(\frac{1}{\epsilon}\right)^2$. This indicates that the required sample size grows significantly for smaller values of α and smaller tolerance values ϵ. We will see this relationship to α play out in our experiments. Nevertheless, this provides a guarantee that is polynomial in all of the relevant parameters. We believe this is the first such guarantee for open category detection.

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 $\hat{F}_a$ then 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 *sufficient*. We say that an anomaly detector is *sufficient* for a problem, if the score CDFs satisfy $F_0(x) \geq F_m(x)$ for all $x \in \mathbb{R}$. Most reasonable anomaly detectors will be sufficient in this sense, since the alien CDF will typically concentrate more mass toward larger anomaly score values compared to F_0.

Corollary 1. *Consider running Algorithm 1 using an upper bound α' on the true α. Under the same assumptions of Theorem 1, if the anomaly detector is sufficient 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 $1 - q - \epsilon$.

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. Our experiments study the impact of using an imprecise α'.

Experiments

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

Synthetic Data Experiments. We first 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 the 3 out of the 9 dimensions are distributed as $N(3,1)$ and with probability 0.6, 4 out of 9 dimensions follow $N(3,1)$.

In each experiment, the nominal data set and the mixture data set are of the same size n, and the mixture data set contains α proportion of anomaly points. The experiments are carried out for $n = 100, 500, 1\,000, 5\,000, 10\,000$ and $\alpha = 0.01, 0.05, 0.10, 0.20, 0.50$. We run the Isolation Forest algorithm to compute $1\,000$ full depth isolation trees on the nominal data. Each tree is grown on a randomly-selected 20% 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 using the full isolation forest. We then apply Algorithm 1 via 10-fold cross validation. We divide the mixture data points at random into 10 groups. For each fold, we estimate $\hat{F}_a$ and τ_a from nine of the 10 groups and then score the mixture points in the held-out fold according to τ_a. The true alarm and false alarm results from each fold are combined to obtain the final output for the entire mixture data set.

For each combination of n and α, we repeated the experiment 100 times. For each run, we record both the alien detection rate, which we call "recall" in the table and the false alarm rate and report the mean and 95% confidence intervals of each. In addition we report the "failure rate", which is the fraction of the 100 runs that did not achieve the desired alien detection rate. In all experiments the desired detection rate is 95%.

We also generate a large nominal data set and a large anomaly data set both of size $20\,000$ in order to obtain "oracle" results for the best possible false alarm rate in each configuration. To get this "oracle", in every experiment, we use the large data set in order to accurately estimate the 5% quantile of the anomaly data and use that threshold to compute an accurate estimate of the false alarm rate using the large anomaly dataset.

Table 1 gives results for our algorithm (with and without isotonization) and the oracle. The "Basic CDF" columns report the results of Algorithm 1 using simple empirical CDF estimates, while the "Iso CDF" uses the isotonized CDF estimates

The first major observation is that for any fixed value of α the performance increase with n as expected. For small sample sizes, the recall generally does not achieve the goal of 95%, which should be expected. Further we see that the required sample size needed to approach achieving the 95% goal increases for smaller values of α. This is predicted by Theorem 1. For $\alpha = 0.5$ we see that the goal is achieved on average as soon as $n = 500$, while for $\alpha = 0.01$ the goal is is not even achieved for $n = 10,000$.

If we consider the false alarm rates, we see that in cases where we are close to achieving the detection-rate goal, the false alarm rate is close to that of the oracle. This indicates that our algorithm has little room to improve in terms of false detection rates. Rather the primary way to improve the false detection rates would be to use more effective anomaly detectors. Note that the failure rates appear to be relatively high and do not approach zero, even for large sample sizes. The reason for this is that the algorithm was run with a goal recall of 95%, and the individuals runs will tend to exhibit small fluctuation on both sides of 95%. Failure rates closer to 0 can be achieved by running the algorithm for recall levels slightly higher than 95%.

We see that the basic CDF method tends to achieve lower recall and lower false alarm rates compared to the isotonized version. This is due to the fact that the smoothing done for the isotonized CDFs tends to result in smaller threshold estimates, which results in higher recall, but also more false alarms.

Benchmark Data Experiments. We now give results for experiments on the five UCI multi-class datasets listed in Table 2. Each data set is constructed by splitting the classes into two groups, nominal and alien. This allows us to gen-

Table 1: Failure rate, recall (i.e. alien detection rate) and false alarm rate from experiments using 9-dimensional normal data, 95%

		Basic CDF				Iso CDF			
		Recall		False Alarm Rate		Recall		False Alarm Rate	
α	n	failrate	recall±CI	FAR±CI	Oracle	failrate	recall±CI	FAR±CI	Oracle
0.01	100	0.40	0.600± 0.098	0.741±0.060	0.939	0.14	0.860±0.069	0.956±0.017	0.939
	500	0.51	0.868±0.030	0.621±0.048	0.825	0.09	0.982±0.011	0.917±0.031	0.825
	1000	0.75	0.859±0.022	0.622±0.050	0.772	0.17	0.975±0.013	0.932±0.025	0.772
	5000	0.76	0.895±0.012	0.487±0.045	0.629	0.22	0.970±0.010	0.856±0.044	0.629
	10000	0.73	0.916±0.008	0.467±0.045	0.586	0.20	0.974±0.008	0.876±0.039	0.586
0.05	100	0.61	0.826±0.034	0.643±0.037	0.742	0.21	0.940±0.026	0.850±0.027	0.742
	500	0.65	0.906±0.015	0.435±0.048	0.477	0.19	0.968±0.012	0.807±0.036	0.477
	1000	0.58	0.924±0.011	0.415±0.044	0.388	0.24	0.970±0.009	0.761±0.046	0.388
	5000	0.69	0.937±0.005	0.280±0.033	0.246	0.24	0.973±0.005	0.682±0.056	0.246
	10000	0.53	0.948±0.004	0.275±0.037	0.208	0.21	0.974±0.006	0.663±0.062	0.208
0.10	100	0.67	0.866 ±0.025	0.476±0.034	0.578	0.35	0.940±0.020	0.703±0.038	0.578
	500	0.64	0.929±0.010	0.364±0.042	0.299	0.23	0.974±0.008	0.707±0.048	0.299
	1000	0.54	0.939±0.007	0.303±0.040	0.232	0.28	0.967±0.008	0.605±0.056	0.232
	5000	0.65	0.947±0.004	0.159±0.025	0.130	0.29	0.968±0.005	0.500±0.061	0.130
	10000	0.55	0.950±0.003	0.137±0.024	0.111	0.35	0.963±0.005	0.363±0.064	0.111
0.20	100	0.45	0.918±0.015	0.400±0.036	0.387	0.21	0.963±0.013	0.603±0.036	0.387
	500	0.63	0.939±0.006	0.222±0.027	0.166	0.33	0.964±0.007	0.417±0.045	0.166
	1000	0.51	0.949±0.005	0.171±0.025	0.121	0.24	0.969±0.005	0.393±0.049	0.121
	5000	0.52	0.950±0.002	0.067±0.004	0.064	0.42	0.956±0.003	0.142±0.031	0.064
	10000	0.59	0.949±0.001	0.051±0.002	0.054	0.52	0.951±0.002	0.076±0.019	0.054
0.50	100	0.56	0.942±0.008	0.149±0.016	0.125	0.33	0.960±0.008	0.233±0.025	0.125
	500	0.49	0.951±0.003	0.057±0.005	0.048	0.39	0.957±0.004	0.097±0.015	0.048
	1000	0.50	0.950±0.002	0.035±0.002	0.034	0.43	0.952±0.002	0.042±0.005	0.034
	5000	0.44	0.950±0.001	0.016±0.001	0.017	0.43	0.950±0.001	0.016±0.001	0.017
	10000	0.40	0.950±0.000	0.013±0.000	0.014	0.40	0.950±0.000	0.013±0.000	0.014

erate nominal data sets S_0 and mixture data sets S_m for any value of α. Since the amount of data in these sets is fixed, we do not vary the value of n as in the previous experiments. We follow an experimental protocol similar to that employed for the synthetic data and give results for three values of α. In addition, for each value of α we run experiments where our algorithm uses different upper bounds $\hat{\alpha} > \alpha$ to observe the impact of imprecise specification.

Overall the results are qualitatively similar to those for the synthetic data. In many configurations the target recall of 95% is achieved on average. The cases where the target are not met are typically for smaller values of α, which our theory predicts has increased data requirements. In the case of Optical Digits the recalls tend to be worse than on other datasets. This is due to the fact that this data set is much smaller than the other datasets, which makes the performance suffer.

We also see that when using an inaccurate upper bound $\hat{\alpha}$ on α the recall and false positive rates tend to increase. This is predicted by the theory and is due to the fact that using an upper bound results in more conservative thresholds. We see, however, that the false alarm rates increase quite rapidly as the gap between $\hat{\alpha}$ and α grows. This indicates that in practice it will be important to use α estimates that are as accurate as possible. Again the isotonic results are significantly more conservative. This suggests that when there is enough data, it is preferable to use the basic CDF approach in order to arrive at better false alarm rates.

Overall these results indicate that using a state-of-the-art anomaly detector with our algorithm results in relatively good performance with respect to recall when the amount of data is sufficient for the particular value of α. The false alarm rates vary significantly across datasets for configurations where the recall constraint is met on average. In general, these rates are controlled by the quality of the anomaly detector, which suggests that the selection of an appropriate detector is critical for achieving practical false alarm rates.

Summary

We have taken a step toward open category detection with guarantees. We found it necessary to assume a mixture model of nominals and aliens with a known mixture rate α to provide guarantees on the alien detection rate. To the best of our knowledge, however, this is the first such guarantee under any similarly restrictive conditions. It is an important open problem to generalize the restrictions under which such guarantees can be derived. Developing methods for reliably estimating upper bounds on α is one avenue for generalization. However, this is a difficult problem both theoretically and empirically.

Our experiments demonstrate that, using a state-of-the-art anomaly detector on benchmark datasets, our algorithm is able to achieve guarantees on the average recall while producing non-trivial false alarm rates. Meeting the guaran-

Table 2: Failure Rate, Recall (i.e. alien detection rate) & False Alarm Rate for UCI Datasets,95%

			Basic CDF			Iso CDF		
				Recall	False Alarm Rate		Recall	False Alarm Rate
Dataset	α	$\hat{\alpha}$	Failrate	recall±CI	FAR±CI	Failrate	recall±CI	FAR±CI
Landsat	0.1	0.100	0.515	0.936 ± 0.007	0.077 ± 0.019	0.240	0.969 ± 0.007	0.400 ± 0.046
	0.1	0.104	0.410	0.951 ± 0.005	0.104 ± 0.022	0.175	0.978 ± 0.005	0.438 ± 0.045
	0.1	0.108	0.245	0.967 ± 0.004	0.148 ± 0.025	0.070	0.987 ± 0.003	0.478 ± 0.043
	0.2	0.200	0.485	0.948 ± 0.004	0.061 ± 0.012	0.295	0.966 ± 0.004	0.267 ± 0.041
	0.2	0.208	0.205	0.967 ± 0.003	0.104 ± 0.017	0.105	0.981 ± 0.003	0.328 ± 0.041
	0.2	0.216	0.035	0.982 ± 0.002	0.184 ± 0.023	0.015	0.991 ± 0.002	0.410 ± 0.041
	0.4	0.400	0.450	0.952 ± 0.002	0.036 ± 0.003	0.395	0.956 ± 0.003	0.084 ± 0.018
	0.4	0.416	0.010	0.974 ± 0.001	0.089 ± 0.010	0.010	0.980 ± 0.002	0.207 ± 0.027
	0.4	0.432	0.000	0.991 ± 0.001	0.249 ± 0.021	0.000	0.994 ± 0.001	0.378 ± 0.031
Opt.digits	0.1	0.100	0.693	0.869 ± 0.015	0.171 ± 0.025	0.382	0.938 ± 0.012	0.484 ± 0.049
	0.1	0.104	0.698	0.889 ± 0.012	0.186 ± 0.026	0.355	0.948 ± 0.010	0.500 ± 0.048
	0.1	0.108	0.656	0.905 ± 0.011	0.213 ± 0.028	0.296	0.955 ± 0.009	0.524 ± 0.047
	0.2	0.200	0.600	0.926 ± 0.008	0.163 ± 0.020	0.355	0.958 ± 0.007	0.396 ± 0.042
	0.2	0.208	0.555	0.940 ± 0.006	0.200 ± 0.024	0.285	0.966 ± 0.006	0.421 ± 0.041
	0.2	0.216	0.395	0.954 ± 0.006	0.245 ± 0.027	0.185	0.975 ± 0.005	0.461 ± 0.040
	0.4	0.400	0.510	0.944 ± 0.005	0.157 ± 0.015	0.365	0.957 ± 0.005	0.284 ± 0.030
	0.4	0.416	0.305	0.964 ± 0.004	0.212 ± 0.019	0.200	0.976 ± 0.004	0.369 ± 0.033
	0.4	0.432	0.100	0.981 ± 0.003	0.303 ± 0.023	0.060	0.989 ± 0.002	0.469 ± 0.033
pageb	0.1	0.100	0.660	0.913 ± 0.012	0.259 ± 0.025	0.460	0.938 ± 0.013	0.413 ± 0.039
	0.1	0.104	0.570	0.929 ± 0.012	0.296 ± 0.027	0.370	0.949 ± 0.012	0.448 ± 0.038
	0.1	0.108	0.450	0.944 ± 0.011	0.337 ± 0.028	0.270	0.961 ± 0.011	0.498 ± 0.037
	0.2	0.200	0.575	0.935 ± 0.010	0.242 ± 0.017	0.425	0.947 ± 0.010	0.341 ± 0.032
	0.2	0.208	0.400	0.952 ± 0.009	0.298 ± 0.020	0.315	0.962 ± 0.010	0.401 ± 0.032
	0.2	0.216	0.195	0.967 ± 0.009	0.376 ± 0.024	0.115	0.976 ± 0.009	0.488 ± 0.031
	0.4	0.400	0.540	0.945 ± 0.004	0.254 ± 0.010	0.445	0.952 ± 0.004	0.288 ± 0.016
	0.4	0.416	0.245	0.963 ± 0.003	0.312 ± 0.012	0.195	0.969 ± 0.003	0.353 ± 0.018
	0.4	0.432	0.050	0.980 ± 0.002	0.390 ± 0.015	0.045	0.984 ± 0.002	0.446 ± 0.021
Shuttle	0.1	0.100	0.550	0.945 ± 0.005	0.171 ± 0.013	0.410	0.958 ± 0.005	0.292 ± 0.033
	0.1	0.104	0.345	0.961 ± 0.004	0.216 ± 0.017	0.210	0.972 ± 0.004	0.361 ± 0.034
	0.1	0.108	0.125	0.978 ± 0.003	0.287 ± 0.023	0.055	0.986 ± 0.003	0.446 ± 0.036
	0.2	0.200	0.505	0.951 ± 0.003	0.159 ± 0.007	0.440	0.954 ± 0.003	0.185 ± 0.015
	0.2	0.208	0.135	0.971 ± 0.003	0.211 ± 0.013	0.130	0.975 ± 0.003	0.266 ± 0.022
	0.2	0.216	0.000	0.989 ± 0.002	0.318 ± 0.020	0.000	0.992 ± 0.002	0.396 ± 0.026
	0.4	0.400	0.485	0.950 ± 0.001	0.147 ± 0.002	0.465	0.951 ± 0.001	0.147 ± 0.002
	0.4	0.416	0.000	0.974 ± 0.001	0.188 ± 0.003	0.000	0.975 ± 0.001	0.191 ± 0.005
	0.4	0.432	0.000	0.995 ± 0.001	0.296 ± 0.009	0.000	0.996 ± 0.001	0.312 ± 0.010
Covertype	0.1	0.100	0.520	0.945 ± 0.007	0.269 ± 0.016	0.480	0.953 ± 0.007	0.171 ± 0.037
	0.1	0.104	0.350	0.960 ± 0.006	0.310 ± 0.020	0.250	0.969 ± 0.006	0.221 ± 0.039
	0.1	0.108	0.160	0.977 ± 0.005	0.377 ± 0.028	0.090	0.984 ± 0.004	0.294 ± 0.040
	0.2	0.200	0.545	0.948 ± 0.003	0.194 ± 0.006	0.510	0.951 ± 0.003	0.211 ± 0.012
	0.2	0.208	0.105	0.970 ± 0.002	0.246 ± 0.011	0.085	0.973 ± 0.002	0.279 ± 0.018
	0.2	0.216	0.000	0.988 ± 0.001	0.345 ± 0.018	0.000	0.991 ± 0.001	0.395 ± 0.021
	0.4	0.400	0.545	0.950 ± 0.002	0.267 ± 0.004	0.520	0.951 ± 0.002	0.268 ± 0.004
	0.4	0.416	0.090	0.969 ± 0.002	0.302 ± 0.005	0.075	0.970 ± 0.002	0.310 ± 0.007
	0.4	0.432	0.000	0.988 ± 0.001	0.384 ± 0.010	0.000	0.989 ± 0.001	0.402 ± 0.012

tees required enough data and the amount of data needed increased for smaller α values. These results agree with our theoretical analysis and indicate the inherent difficult of open category detection in the small α setting. 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 threat, there is little recourse other than to collect more data or allow for higher false positive rates.

Acknowledgements

This research was supported by a gift from Huawei, Inc., and grants from the Future of Life Institute and the NSF Grant 1514550. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the au-

thor(s) and do not necessarily reflect the views of the sponsors.

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

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

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 $\hat{\tau}_q\le\tau_{q+\epsilon}$. To achieve this, we want to have

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

We have

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

Making use of (2), when

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

we will have

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

In this case we will have

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

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

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

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

$$F_a(\hat{\tau}_q)\le\hat{F}_a(\hat{\tau}_q)+\epsilon\le q+\epsilon,$$

and thus

$$\hat{\tau}_q\le\tau_{q+\epsilon}.$$

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

Proof for Corollary 1

If $\alpha'\ge\alpha$, 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

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

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

But

$$F'_a(x)-F_a(x)=\frac{(\alpha-\alpha')(F_m(x)-F_0(x))}{\alpha\alpha'}\ge 0,\forall x\in\mathbb{R}.$$

and because of this, if we let $\tau'_{q+\epsilon}$ denote the $q+\epsilon$ quantile of F'_a, we will have $\tau'_{q+\epsilon}\le\tau_{q+\epsilon}$. 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 $1-\delta$, $\hat{\tau}'_q\le\tau'_{q+\epsilon}\le\tau_{q+\epsilon}$.□

References

Barlow, R., and Brunk, H. 1972. Problem The Isotonic Regression and Its Dual. *Journal of the American Statistical Association* 67(337):140–147.

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

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

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

Da, Q.; Yu, Y.; and Zhou, Z.-H. 2014. Learning with augmented class by exploiting unlabeled data. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, 1760–1766. AAAI Press.

Emmott, A. F.; Das, S.; Dietterich, T.; Fern, A.; and Wong, W.-K. 2013. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, 16–21. ACM.

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

Hendrycks, D., and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR* abs/1610.02136.

Jin, H.; Liu, Q.; and Lu, H. 2004. Face detection using one-class-based support vectors. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, 457–462.

Liang, S.; Li, Y.; and Srikant, R. 2017. Principled detection of out-of-distribution examples in neural networks. *CoRR* abs/1706.02690.

Liu, F. T.; Ting, K. M.; and Zhou, Z.-H. 2008. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, 413–422. IEEE.

Lytle, D. A.; Martínez-Muñoz, G.; Zhang, W.; Larios, N.; Shapiro, L.; Paasch, R.; Moldenke, A.; Mortensen, E. N.; Todorovic, S.; and Dietterich, T. G. 2010. Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society* 29(3):867–874.

Manevitz, L. M., and Yousef, M. 2002. One-class svms for document classification. *J. Mach. Learn. Res.* 2:139–154.

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

Mendes Júnior, P. R.; de Souza, R. M.; Werneck, R. d. O.; Stein, B. V.; Pazinato, D. V.; de Almeida, W. R.; Penatti, O. A. B.; Torres, R. d. S.; and Rocha, A. 2017. Nearest neighbors distance ratio open-set classifier. *Machine Learning* 106(3):359–386.

Pietraszek, T. 2005. Optimizing abstaining classifiers using roc analysis. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, 665–672. New York, NY, USA: ACM.

Pritsos, D. A., and Stamatatos, E. 2013. *Open-Set Classification for Automated Genre Identification*. Berlin, Heidelberg: Springer Berlin Heidelberg. 207–217.

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

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

Schölkopf, B.; Platt, J. C.; Shawe-Taylor, J. C.; Smola, A. J.; and Williamson, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural Comput.* 13(7):1443–1471.

Shu, L.; Xu, H.; and Liu, B. 2017. DOC: deep open classification of text documents. *CoRR* abs/1709.08716.

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

Wegkamp, M. H. 2007. Lasso type classifiers with a reject option.

Wu, M., and Ye, J. 2009. 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.

Zhou, X. S., and Huang, T. S. 2003. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems* 8(6):536–544.