Open Category Detection with PAC Guarantees

Si Liu [*1] **Risheek Garrepalli** [*2] **Thomas G. Dietterich** [2] **Alan Fern** [2] **Dan Hendrycks** [3]

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. 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 with PAC-style guarantees on the alien detection rate, while aiming to minimize false alarms. Empirical results on synthetic and standard benchmark datasets demonstrate the regimes in which the algorithm can be effective and provide a baseline for further advancements.

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" data, 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 where a classifier is trained 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 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 scientific conclusions.

The problem of open category detection is to detect such alien examples 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, no existing algorithm provides 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 accuracy 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, say, 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 two training sets: a clean training dataset involving a finite set of cate-

[*] Equal contribution [1] Department of Statistics, Oregon State University, Oregon, USA [2] School of EECS, Oregon State University, Oregon, USA [3] University of California, Berkeley, California, USA. Correspondence to: Si Liu <lius2@oregonstate.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

gories and a contaminated dataset that contains a fraction α of aliens. 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 uses 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 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 clean and contaminated training data in order to derive an alarm threshold that achieves the desired guarantee on the alien detection rate on new test queries. 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.

We carry out experiments[1] on synthetic and benchmark datasets using a state-of-the-art anomaly detector, the Isolation Forest (Liu et al., 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.

2. 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 & Huang, 2003), which has been partly addressed by later work (Manevitz & Yousef, 2002; Wu & Ye, 2009; Jin et al., 2004; Cevikalp & Triggs, 2012). OCSVMs have been employed in a multi-class setting similar to open category detection (Heflin et al., 2012; Pritsos & 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 & Duin, 2008; Pietraszek, 2005; Geifman & 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 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 et al., 2014). Additional work has addressed open category detection by tuning the decision boundary based on unlabeled data which contains data from novel categories (Da et al., 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 recent interest in open category detection for deep neural networks applied to vision and text classification (Bendale & 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. 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 & 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 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 containing 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" dataset S_m that contains n points drawn i.i.d. from a mixture dis-

[1] Code for reproducing our experiments can be found at https://github.com/liusi2019/ocd.

tribution 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.

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 datasets, our problem is to label test instances from D_m as either "alien" or "nominal". In particular, we wish to achieve a specified *alien detection 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* 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. Intuitively, the anomaly scores order 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 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 threshold that provides a guarantee on the alien detection rate.

4. Algorithms for Open Category Detection

In order to obtain theoretical guarantees, 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.

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

```
1: Get 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 & 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 supplementary materials) 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 on the uniform convergence of empirical CDF functions (Massart, 1990). To use this result, we make the reasonable technical assumption that the nominal and alien CDFs, F_0 and F_a, are continuous with convex support. 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 datasets 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 $O(\frac{1}{\epsilon^2 \alpha^2 \sqrt{\delta}})$. 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 gives us flexibility in setting ϵ and q (the algorithm input) to achieve a guarantee of $1-\eta$. The ϵ parameter controls a trade-off between sample size and false alarm rate. To minimize the false alarm rate, we want to make q large (to obtain a larger threshold), so we want to set q close to η. But, as $q \to \eta$, $\epsilon \to 0$, and $n \to \infty$. To minimize the sample size n, we want to make q as small as possible, because that allows ϵ to be larger and hence n becomes smaller. The optimal setting of ϵ depends on how the false alarm rate grows with τ_q, which in turn depends on the relative shape of F_0 and F_a. In a real safety application, we can estimate these from S_0 and S_m and choose an appropriate q value.

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.

Corollary 1. *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. Experiments

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 datasets? 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 (Liu et al., 2008), which has been demonstrated to be a state-of-the-art detector in recent empirical studies (Emmott et al., 2013). In the Supplementary Materials we show similar results with the LODA anomaly detector (Pevný, 2015).

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.

In each experiment, the nominal dataset and the mixture dataset are of the same size n, and the mixture dataset contains a proportion α of anomaly points. We fixed the target quantile to be $q = 0.05$. The experiments are

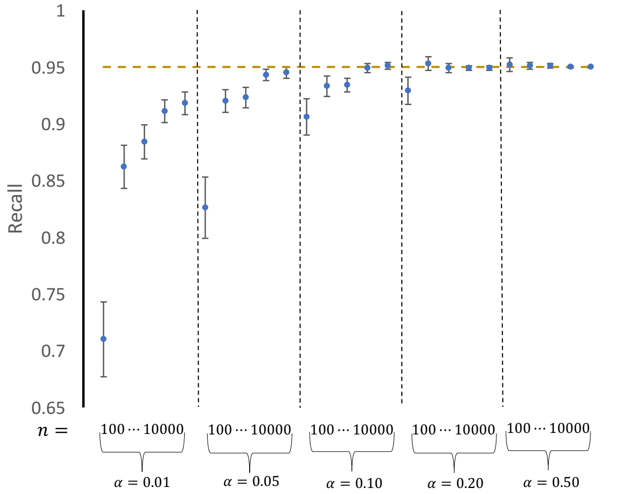

Figure 1. Comparison of recall achieved by $\hat{\tau}_q$ compared to oracle recall of 0.95. Error bars are 95% confidence intervals. Settings of n and α increase from left to right starting with $\alpha = 0.01$ and $n \in \{100, 500, 1\text{K}, 5\text{K}, 10\text{K}\}$ up to $\alpha = 0.5$ and $n =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

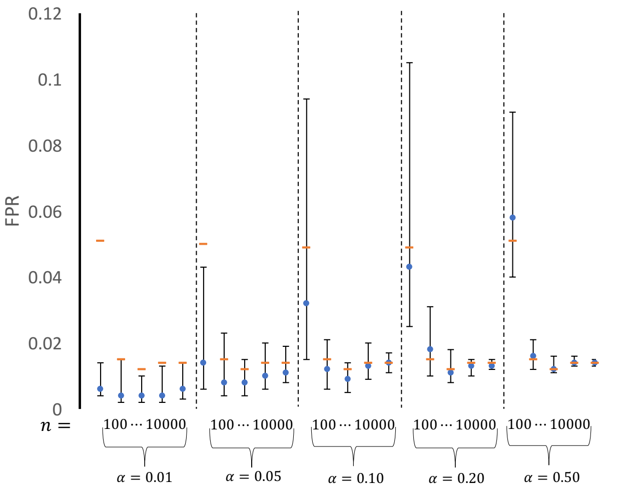

Figure 2. Comparison of oracle FPR to the FPR achieved by $\hat{\tau}_q$. Error bars span from the 25th to 75th percentile with the blue dot marking the median of the 100 trials. Orange markers indicate the oracle FPR. Settings of n and α increase from left to right starting with $\alpha = 0.01$ and $n \in \{100, 500, 1\text{K}, 5\text{K}, 10\text{K}\}$ up to $\alpha = 0.5$ and $n = 10\text{K}$.

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 datasets G_0 and G_a, with G_0 being a pure nominal dataset, G_a being a pure alien dataset, 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. 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 1 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 2 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 3 shows a

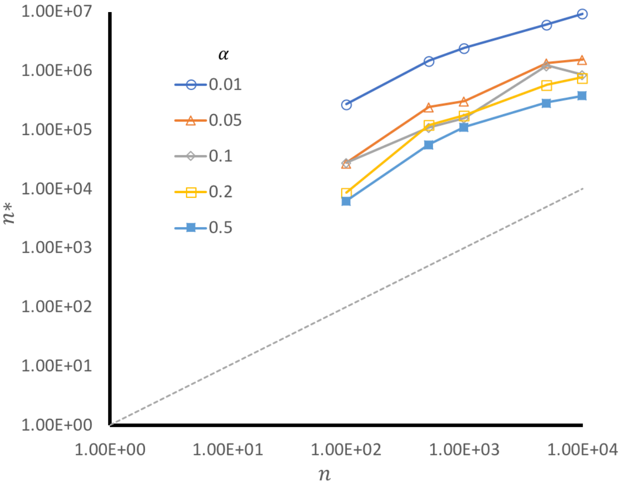

Figure 3. The log sample size n^* required by Theorem 1 in order to guarantee the actual observed recall versus the log actual sample size n.

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.

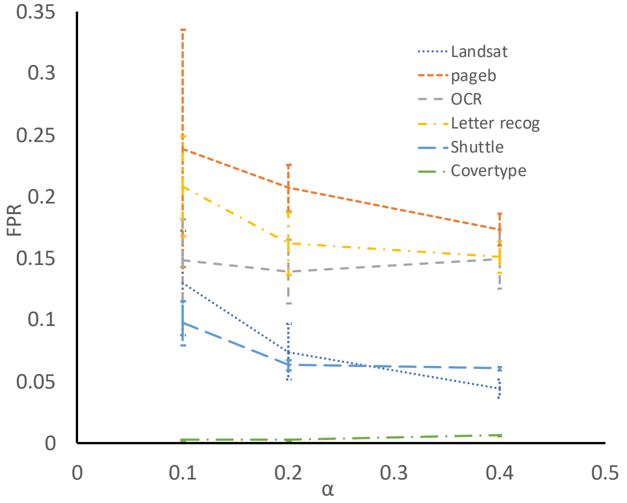

Figure 4. False positive rates on six UCI datasets as a function of α ($q = 0.05$, $\delta = 0.05$).

Benchmark Data Experiments. To address our third and fourth questions, we performed experiments on six UCI multiclass datasets: Landsat, Opt.digits, pageb, Shuttle, Covertype and MNIST. In addition to these, we provide results for the Tiny ImageNet dataset. In each multiclass dataset, 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 datasets are MNIST(1,3,7), Landsat(1,7), OCR(1,3,4,5,7), pageb(1,5), Letter recognition(1,3), and Shuttle(1,4). We generated nominal and mixture datasets for various values of α. The value of n for each dataset is 1532 for Landsat, 568 for Opt.digits, 4912 for pageb, 5000 for Shuttle, 13,624 for Covertype, 11,154 for MNIST, and 10,000 for Tiny ImageNet. Because we cannot create datasets with large n, we cannot measure the true value of τ_q.

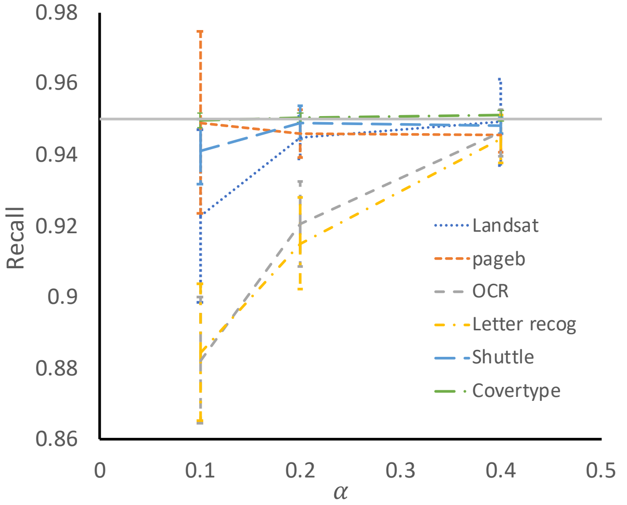

Figure 5. Recall rates on six UCI datasets as a function of α ($q = 0.05$, $\delta = 0.05$)

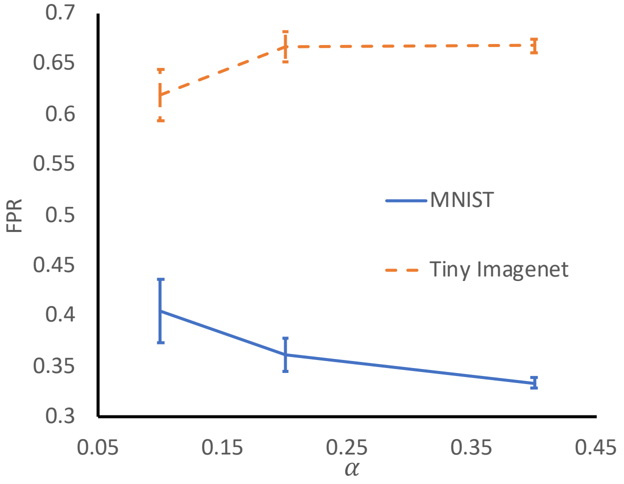

Figure 6. False positive rates on two image datasets as a function of α ($q = 0.05$, $\delta = 0.05$).

After computing the anomaly scores for both nominal and mixture datasets, 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 a baseline method (Hendrycks & Gimpel, 2017).

To answer Q3, Figures 4 and 6 plot the false positive rate as a function of α for the UCI and vision datasets, respectively. We see that the FPR ranges from 3.6% to 26.9% on UCI depending on the dataset and the level of α. The vision datasets have higher FPR, especially MNIST, which has a large number of alien classes that are not distinguished well by the anomaly detector. The FPR depends primarily on the domain, because the key issue is how well the anomaly detector distinguishes between nominal and alien examples. The false alarm rate 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 datasets having different (or unknown) values of α.

Figures 5 and 7 plot the recall rate as a function of α for the UCI and vision datasets. We set $q = 0.05$ in these experiments. Theorem 1 only guarantees a recall of $1-q-\epsilon$,

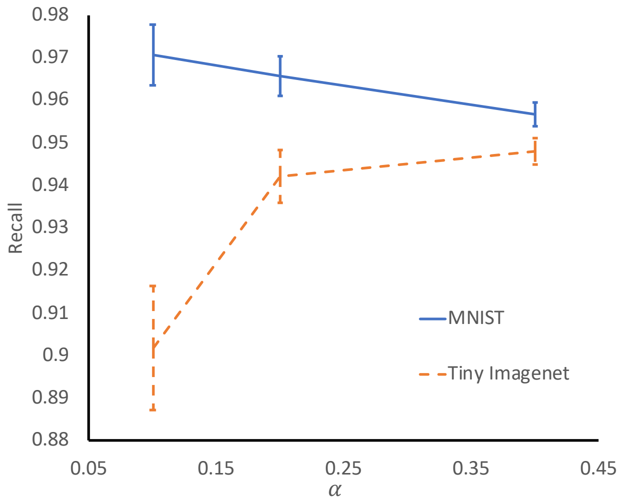

Figure 7. Recall rates on two image datasets as a function of α ($q = 0.05, \delta = 0.05$).

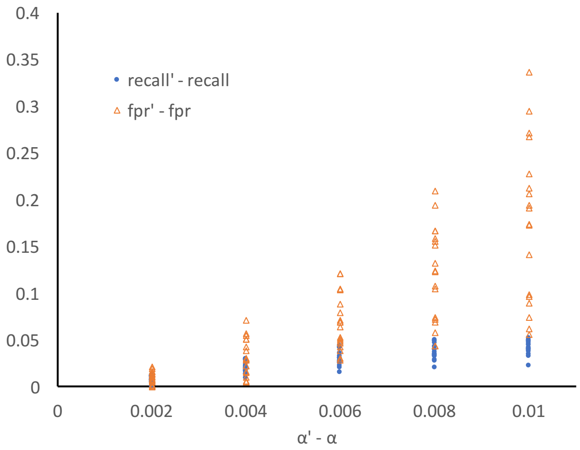

Figure 8. Change in recall and false positive rate as a function of $\alpha' - \alpha$ for six UCI datasets; $\alpha \in \{0.1, 0.2, 0.4\}$

where ϵ depends on n. Hence, it is nice to see that for three of the domains (Shuttle, Covertype, and Landsat) in UCI and for both vision datasets, 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 datasets when $\alpha = 0.1$. This shows that in applications such as fraud detection, where α may be very small, the mixture dataset S_m needs to be very large.

To answer Q4 regarding the impact of using an incorrect value $\alpha' > \alpha$, we repeated these experiments with $\alpha' = \alpha + \xi$, for $\xi \in \{0.002, 0.004, 0.006, 0.008, 0.010\}$. Figure 8 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 dataset, the change in Recall and the change in FPR. We observe that the recall increases slightly (in the range from 0.01 to 0.05). However, the false positive rate increases by much larger amounts (from 0.01 to 0.336). This demonstrates that it is very important to determine the value of α accurately.

7. Summary

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. Nonetheless, our experiments show that our algorithm gives good recall performance and non-trivial false alarm rates on datasets 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 datasets: a clean dataset that contains no aliens and an (unlabeled) contaminated dataset that contains a known fraction α of aliens. An important theoretical problem for future research is to develop a method that can estimate a tight upper bound on $\hat{\alpha} > \alpha$. We believe this is possible, but we have not yet found a method that guarantees that $\hat{\alpha} > \alpha$.

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.

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

A. 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

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

holds without any restriction on λ. Making use of this, and assuming we use the same sample size n for both the mixture dataset 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 $\hat{\tau}_q \leq \tau_{q+\epsilon}$. 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 (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

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

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

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 $1-\delta$,

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

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

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

and thus

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

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

B. Proof for Corollary 1

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

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

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

But

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Mendes Júnior, Pedro R., de Souza, Roberto M., Werneck, Rafael de O., Stein, Bernardo V., Pazinato, Daniel V., de Almeida, Waldir R., Penatti, Otávio A. B., Torres, Ricardo da S., and Rocha, Anderson. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106 (3):359–386, Mar 2017. ISSN 1573-0565.

Pevný, Tomáš. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, (November 2014), 2015.

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

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

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

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

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

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

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

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

Wu, M. and Ye, J. 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.

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