Finite Sample Complexity of Rare Pattern Anomaly Detection

Md Amran Siddiqui and **Alan Fern** and **Thomas G. Dietterich** and **Shubhomoy Das**
School of EECS
Oregon State University
{siddiqmd, afern, tgd, dassh}@eecs.oregonstate.edu

Abstract

Anomaly detection is a fundamental problem for which a wide variety of algorithms have been developed. However, compared to supervised learning, there has been very little work aimed at understanding the sample complexity of anomaly detection. In this paper, we take a step in this direction by introducing a Probably Approximately Correct (PAC) framework for anomaly detection based on the identification of rare patterns. In analogy with the PAC framework for supervised learning, we develop sample complexity results that relate the complexity of the pattern space to the data requirements needed for PAC guarantees. We instantiate the general result for a number of pattern spaces, some of which are implicit in current state-of-the-art anomaly detectors. Finally, we design a new simple anomaly detection algorithm motivated by our analysis and show experimentally on several benchmark problems that it is competitive with a state-of-the-art detector using the same pattern space.

1 INTRODUCTION

The problem of (unsupervised) anomaly detection is to identify anomalies in (unlabeled) data, where an anomaly is a data point that is generated by a process that is distinct from the process that generates "normal" points. This problem arises in a large number of applications, from security to data cleaning, and there have been many approaches proposed in the literature [4, 8]. While most applications seek to identify semantically-interesting anomalies, it is typically not possible to predefine a functional notion of semantic interestingness. Instead, the vast majority of anomaly detectors use a surrogate measure of interestingness. For example, a point may be interesting if it is a statistical outlier or if it is far away from other data points. The performance of a given detector in a domain depends on how well it can optimize the statistical measure and on how well that measure aligns with the behavior of anomalies in the application domain.

For moderately high-dimensional data, all data points become far apart from each other, so they are all statistical outliers in a sense. This suggests that anomaly detection by identifying outliers or distant points should perform poorly and degrade to random selection as the dimensionality grows. Empirical results, however, have shown that state-of-the-art anomaly detectors often perform quite well [7] even for high-dimensional data. Further, these detectors tend to reach their peak performance with a relatively small amount of training data compared to what might be expected based on the dimensionality. The primary goal of this paper is to move toward an understanding of these empirical observations by analyzing the sample complexity of a certain class of anomaly detectors.

The sample complexity of supervised learning has been widely studied and is quite well understood via the framework of Probably Approximately Correct (PAC) learning. However, this is not the case for anomaly detection, where virtually all published work has focused on algorithms with good empirical performance (with additional attention to computational speed, especially on big data sets). A key step in the development of PAC learning theory was to formalize the notion of a hypothesis space and to quantify the relationship between the complexity of this space and the amount of training data required to identify a good hypothesis in the space. In this paper, we follow a similar approach. Our framework is motivated by the observation that many state-of-the-art anomaly detectors can be viewed as monitoring the probabilities of certain "patterns" in the data, where a "pattern" is a subset (typically closed and compact) of the feature space. Outliers are then identified based on measures of those probabilities, where points are ranked as more anomalous if they satisfy lower-probability patterns. For example, the highly-competitive anomaly detection algorithm, Isolation Forest [10], finds outliers by monitoring probabilities in the pattern space of axis-aligned hyper-rectangles. Section 5 provides additional ex-

amples of the pattern spaces underlying a number of other state-of-the-art detectors. In our analysis, a "pattern" will play the same role as a "hypothesis" in PAC learning, and the pattern space complexity will determine the number of training examples required for high accuracy.

A second key step in the development of PAC theory was to relax the goal of finding the best possible hypothesis. Similarly, we will introduce an error parameter ϵ that determines how accurately the algorithm must estimate the probabilities of the patterns in the pattern space. We will then show that the required sample size scales polynomially in $1/\epsilon$ (as well as in several other parameters).

We call our formulation Rare Pattern Anomaly Detection (RPAD), and an algorithm that provides PAC guarantees will be referred to as a PAC-RPAD algorithm. We prove sample complexity results for any algorithm within the RPAD framework. The framework captures the qualitative essence of many anomaly detection algorithms. Note that we focus exclusively on sample complexity. Experience with the supervised PAC analysis has shown that sample complexity results often give more insight than computational complexity results. Indeed, many computational problems in PAC learning are NP-Hard and yet practical approximate algorithms are known. Similarly, we expect that some of the computational PAC-RPAD problems will also be NP-Hard, but existing algorithms, such as Isolation Forest, already provide practical approximate solutions.

Prior work on one-class SVMs [12] and learning minimum volume sets [13] have also provided sample complexity analysis relevant to anomaly detection. These approaches, however, are fundamentally different than RPAD-style approaches. In particular, these approaches focus on finding a common region/pattern in the input space that capture the normal points. In contrast, our RPAD framework is based on finding rare regions/patterns for directly extracting anomaly characteristics. Anomaly detection benchmarking studies [7] have shown that RPAD-style approaches tend to significantly outperform "common pattern" approaches such as one-class SVM. Our work is the first to analyze the sample complexity of the former approach.

The main contributions of this paper are as follows. First, we present a formal framework for RPAD (Section 2), which leads to the definition of PAC-RPAD (Section 3). Second, we specify a simple generic algorithm, RAREPATTERNDETECT, based on finding rare patterns. We derive sample complexity results for both finite pattern spaces and uncountable spaces of bounded complexity (Section 4). Third, we give a number of applications of the theory to pattern spaces that underly a number of state-of-the-art anomaly detection algorithms (Section 5). This, in part, helps explain why such algorithms consistently perform much better than random in high-dimensional data. Fourth, we measure learning curves on several benchmarks and for pattern spaces of varying complexity for RAREPATTERNDETECT and another state-of-the-art anomaly detector over the same spaces (Section 6). The results show that the RPAD-based algorithm can be competitive with the state-of-the-art and that the detectors' performances converge for surprisingly small training sets.

2 RARE PATTERN ANOMALY DETECTION

We consider anomaly detection over a space of possible data points $\mathcal{X} \subseteq \mathcal{R}^d$, which may be finite or infinite. Data from this space is generated according to an unknown probability density function $\mathcal{P}$ over $\mathcal{X}$. A common assumption in anomaly detection is that $\mathcal{P}$ is a mixture of a normal component, which generates normal data points, and an anomaly component, which generates anomalous data points. Further, it is typically assumed that there is a much higher probability of generating normal data than anomalies. This set of assumptions motivates one approach to anomaly detection, which we call *Anomaly Detection via Outlier Detection*. The idea is to estimate, for each query point x, the density $\mathcal{P}(x)$ based on an (unlabeled) training sample of the data and assign an anomaly score to x proportional to $-\log \mathcal{P}(x)$, the "surprise" of x.

There are many problems with this approach. First, the probability density may not be smooth in the neighborhood of x, so that $\mathcal{P}(x)$ could be very large and yet be surrounded by a region of low or zero density (or vice versa). Second, even under smoothness assumptions, density estimation is very difficult. For example, the integrated squared error of kernel density estimation in d-dimensional space (for a second-order kernel, such as a Gaussian kernel) converges to zero at a rate of $O(N^{-4/(4+d)})$ [14]. It follows that the sample size N required to achieve a target accuracy grows exponentially in the dimension d.

In this paper, we consider an alternative anomaly detection framework, which we will refer to as *Rare Pattern Anomaly Detection (RPAD)*. Informally, the main idea is to judge a point as anomalous if it exhibits a property, or pattern, that is rarely exhibited in data generated by $\mathcal{P}$. For example, in a computer security application, a detector may monitor various behavior patterns associated with processes accessing files. A process that exhibits an access behavior pattern that has been rarely seen would be considered anomalous. One attractive feature of the RPAD framework is that the notion of anomaly is grounded in the estimation of pattern probabilities, rather than point densities. Pattern probability estimation is quite well understood compared to density estimation. A second attractive feature of the RPAD framework is that each detected anomaly comes with an explanation of why it was considered anomalous. Specifically, the explanation can report the rare patterns that the

anomaly satisfies. Explanation methods have been developed for density-estimation approaches (e.g. [15]), but they are less directly tied to the anomaly detection criterion.

Formally, a *pattern* h is a binary function over $\mathcal{X}$. A *pattern space* $\mathcal{H}$ is a set of patterns, which may be finite or infinite. As an example, if $\mathcal{X}$ is a finite space of n-dimensional bit vectors, a corresponding finite pattern space could be all conjunctions of length up to k. As another example, let $\mathcal{X} = [0,1]^n$, the n-dimensional unit cube, and consider the uncountable pattern space of all axis-aligned k-dimensional hyper-rectangles in this cube. In this case, each pattern h is a hyper-rectangle, such that $h(x)$ is true if x falls inside it. The choice of pattern space is an important consideration in the RPAD framework, since, in large part, this choice controls the semantic types of anomalies that will be detected.

Each pattern $h \in \mathcal{H}$ has a probability $P(h) = \Pr(\{x : h(x) = 1\})$ of being satisfied by data points generated according to $\mathcal{P}$. It will be useful to specify the set of all patterns in $\mathcal{H}$ that a point x satisfies, which we will denote by $\mathcal{H}[x] = \{h \in \mathcal{H} : h(x) = 1\}$. One approach to RPAD is to declare x to be anomalous if there is a pattern $h \in \mathcal{H}[x]$, such that $P(h) \leq \tau$. This approach is sensible when all patterns are approximately of the same "complexity". However, when $\mathcal{H}$ contains a mix of simple and complex patterns, this approach can be problematic. In particular, more complex patterns are inherently less likely to be satisfied by data points than simpler patterns, which makes choosing a single threshold τ difficult. For this reason, we introduced the *normalized pattern probability* $f(h) = P(h)/U(h)$, where $U(h)$ is the probability of h being satisfied according to a *reference density* U over $\mathcal{X}$. When $\mathcal{X}$ is bounded, we typically take U to be a uniform density function. Thus, a small value of $f(h)$ indicates that under $\mathcal{P}$, h is significantly more rare than it would be by chance, which provides a better-calibrated notion of rareness compared to only considering $\mathcal{P}(h)$. If $\mathcal{X}$ is unbounded, then an appropriate maximum entropy distribution (e.g., Gaussian) can be chosen.

We can now define the notion of anomaly under the RPAD framework. We say that a pattern h is *τ-rare* if $f(h) \leq \tau$, where τ is a detection threshold specified by the user. A data point x is a *τ-outlier* if there exists a τ-rare h in $\mathcal{H}[x]$ and otherwise x is said to be τ-common. Given τ and a stream of data drawn from $\mathcal{P}$, an optimal detector within the RPAD framework should detect all τ-outlier points and reject all τ-common points. That is, we want to detect any point that satisfies a τ-rare pattern and otherwise reject. An anomaly detector will make its decisions based on some finite amount of sampled data, and we expect that the performance should improve as the amount of data grows. Further, in analogy to supervised learning, we would expect that the amount of data required to achieve a certain level of performance should increase as the complexity of the pattern space increases. We now introduce a formal framework for making these intuitions more precise.

3 PROBABLY APPROXIMATELY CORRECT FRAMEWORK

To address the sample complexity of RPAD, we consider a learning protocol that makes the notion of training data explicit. The protocol first draws a training data set $\mathcal{D}$ of N i.i.d. data points from $\mathcal{P}$. An anomaly detector is provided with $\mathcal{D}$ along with a test instance x that may or may not be drawn from $\mathcal{P}$. The anomaly detector then outputs "detect" if the instance is considered to be an anomaly or "reject" otherwise. Note that the output of a detector is a random variable due to the randomness of $\mathcal{D}$ and any randomization in the algorithm itself. This protocol models a common use case in many applications of anomaly detection. For example, in a computer security application, data from normal system operation will typically be collected and provided to an anomaly detector before it is activated.

The *ideal* correctness criterion requires that the test instance x be detected if it is a τ-outlier and rejected otherwise. However, as we discussed above, this notion of correctness is too strict for the purpose of sample complexity analysis. In particular, such a criterion requires distinguishing between pattern probabilities that fall arbitrarily close to each side of the detection threshold τ, which can require arbitrarily large training samples. For this reason, we relax the correctness criterion by introducing a *tolerance parameter* $\epsilon > 0$. The detector is said to be approximately correct at level ϵ if it detects all τ-rare points and rejects all $(\tau+\epsilon)$-common points. For test points that are neither τ-rare nor $(\tau+\epsilon)$-common, the detector output can be arbitrary. The value of ϵ controls the false positive rate of the detector, where smaller values of ϵ will result in fewer false positives relative to the detection threshold.

We now define the PAC learning objective for RPAD. A detection algorithm will be considered PAC-RPAD if with high-probability over draws of the training data it produces an approximately correct output for any $x \in \mathcal{X}$.

Definition 1. *(PAC-RPAD) Let $\mathcal{A}$ be a detection algorithm over pattern space $\mathcal{H}$ with input parameters $0 < \delta < 1$, $0 < \tau$, $0 < \epsilon$, and the ability to draw a training set $\mathcal{D}$ of any size N from $\mathcal{P}$. $\mathcal{A}$ is a PAC-RPAD algorithm if for any $\mathcal{P}$ and any τ, with probability at least $1-\delta$ (over draws of $\mathcal{D}$), $\mathcal{A}$ detects all τ-outliers and rejects all $(\tau+\epsilon)$-commons.*

The *sample complexity* of a PAC-RPAD algorithm for $\mathcal{H}$ is a function of the inputs $N(\delta, \epsilon)$ that specifies the number of training examples to draw. We expect that the sample complexity will increase as the complexity of $\mathcal{H}$ increase, as the dimensionality d of points increases, and as the failure probability δ decreases. Further, we expect that the sample complexity will increase for smaller values of the

tolerance parameter ϵ, since this controls the difficultly of distinguishing between τ-rare and $(\tau+\epsilon)$-common data points. Accordingly we say that a PAC-RPAD algorithm is *sample efficient* if its sample complexity is polynomial in d, $\frac{1}{\delta}$, and $\frac{1}{\epsilon}$.

4 FINITE SAMPLE COMPLEXITY OF RPAD

We now consider a very simple algorithm, called RAREPATTERNDETECT, which will be shown to be a sample efficient PAC-RPAD algorithm for bounded complexity pattern spaces. The algorithm is given in Table 1 and first draws a training set $\mathcal{D}$ of size $N(\delta,\epsilon)$. Here $N(\delta,\epsilon)$ will depend on the pattern space complexity and will be specified later in this section. The training set is used to estimate the normalized pattern probabilities given by

$$\widehat{f}(h) = \frac{1}{|D|\cdot U(h)}|\{x\in\mathcal{D} : h(x)=1\}|.$$

Here, we assume that $U(h)$ can be computed analytically or at least closely approximated. For example, when U is uniform over a bounded space, $U(h)$ is proportional to the volume of h.

After drawing the training set, RAREPATTERNDETECT specifies a decision rule that detects any x as anomalous if and only if it satisfies a pattern with estimated frequency less than or equal $\tau+\epsilon/2$. This test is done using the subroutine HASRAREPATTERN$(x,\mathcal{D},\mathcal{H},\mu)$, which returns true if there exists a pattern h in $\mathcal{H}[x]$ such that $\widehat{f}(h)\leq\mu$. For the purposes of sample complexity analysis, we will assume an oracle for HASRAREPATTERN. For sufficiently complex pattern spaces, the problem addressed by HASRAREPATTERN will be computational hard. Thus, in practice, a heuristic approximation will be needed, for example, based on techniques developed in the rare pattern mining literature. In practice, we are often interested in having an anomaly detector return an anomaly ranking over multiple test data points. In this case, the algorithm can rank a data point x based on a score equal to the minimum normalized frequency of any pattern that it satisfies, that is, score$(x) = \min\{\widehat{f}(h) : h\in\mathcal{H}[x]\}$. It remains to specify $N(\delta,\epsilon)$ in order to ensure that RAREPATTERNDETECT is PAC-RPAD. Below we do this for two cases: 1) finite pattern spaces, and 2) pattern spaces with bounded VC-dimension. Later, in Section 5 we will instantiate these results for specific pattern spaces that underly several existing anomaly detection algorithms.

4.1 SAMPLE COMPLEXITY FOR FINITE $\mathcal{H}$

For finite pattern spaces, it is relatively straightforward to show that as long as $\log|\mathcal{H}|$ is polynomial in d then RAREPATTERNDETECT is a sample efficient PAC-RPAD algorithm.

Table 1: RAREPATTERNDETECT Algorithm

Input: δ, τ, ϵ

1. Draw a training set $\mathcal{D}$ of $N(\delta,\epsilon)$ instances from $\mathcal{P}$.
2. **Decision Rule for any x:**
   If HASRAREPATTERN$(x,\mathcal{D},\mathcal{H},\tau+\epsilon/2)$ then return "detect", otherwise return "reject".

HASRAREPATTERN$(x,\mathcal{D},\mathcal{H},\mu)$
$:= \{h\in\mathcal{H}[x] : \widehat{f}(h)\leq\mu\}\neq\emptyset$

Theorem 1. *For any finite pattern space $\mathcal{H}$,* RAREPATTERNDETECT *is PAC-RPAD with sample complexity* $N(\delta,\epsilon) = O\left(\frac{1}{\epsilon^2}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)\right)$.

Proof. Suppose, X is a Bernoulli random variable with parameter $P(h)$ for a pattern h, i.e., $E[X]=P(h)$. Let, $Y=\frac{X}{U(h)}$, hence, Y is a random variable with $E[Y]=\frac{E[X]}{U(h)}=\frac{P(h)}{U(h)}=f(h)$. We also observe that the maximum value of Y is $\frac{1}{U(h)}$. Given N samples $x_1,x_2,...,x_N$, each $x_i\sim\mathcal{P}$, we estimate $\widehat{f}(h)=\frac{1}{N}\sum_{i=1}^{N}\frac{\mathcal{I}[x_i\in h]}{U(h)}$. We seek a confidence interval $[L(h),R(h)]$ for $f(h)$ that is narrow enough that it does not simultaneously contain both τ and $\tau+\epsilon$. The reason is that if $L(h)>\tau$, then with probability $1-\delta$, $f(h)>\tau$, so h is not a τ-rare pattern. If $R(h)<\tau+\epsilon$, then with probability $1-\delta$, h is not a $\tau+\epsilon$-common pattern, so it is safe to treat it as τ-rare. Hence, the confidence interval should be $[\widehat{f}(h)-\epsilon/2, \widehat{f}(h)+\epsilon/2]$, and its "half width" is $\epsilon/2$. So, we want to bound (by δ) the probability that $\widehat{f}(h)$ is more than $\frac{\epsilon}{2}$ away from its true value $f(h)$. Now, using the Hoeffding bound we have

$$\begin{aligned}
&P\left(|E_{\mathcal{P}}[\widehat{f}(h)]-\widehat{f}(h)|>\frac{\epsilon}{2}\right)\\
\iff\ &P\left(\left|f(h)-\frac{1}{N}\sum_{i=1}^{N}\frac{\mathcal{I}[x_i\in h]}{U(h)}\right|>\frac{\epsilon}{2}\right)\\
&\leq 2\exp\left(-\frac{\epsilon^2}{2}U(h)^2N\right).
\end{aligned}$$

Since $\mathcal{H}$ is finite, we can bound the above probability for all $h\in\mathcal{H}$ using the union bound: $2|\mathcal{H}|\exp\left(-\frac{\epsilon^2}{2}U_{min}^2N\right)$, where, $U_{min}=\min_{h\in\mathcal{H}}U(h)$. We want this quantity to be less or equal to δ:

$$\begin{aligned}
&2|\mathcal{H}|\exp\left(-\frac{\epsilon^2}{2}U_{min}^2N\right)\leq\delta\\
\implies\ &N\geq\frac{2}{\epsilon^2}\frac{1}{U_{min}^2}\log\frac{2|\mathcal{H}|}{\delta}.
\end{aligned}$$

Hence, the sample complexity is $O\left(\frac{1}{\epsilon^2}\left(\log|\mathcal{H}|+\log\frac{1}{\delta}\right)\right)$. □

4.2 SAMPLE COMPLEXITY FOR INFINITE $\mathcal{H}$

When the sample space $\mathcal{X}$ is continuous, it is typically the case that the corresponding pattern space will be infinite and hence not covered by the above result. As is standard in supervised learning, we will characterize the complexity of infinite pattern spaces via the VC-dimension [17], which we denote by $\mathcal{V}_\mathcal{H}$. The VC-dimension of $\mathcal{H}$ is equal to the maximum number of points that can be shattered by patterns in $\mathcal{H}$. Here a set of points D is shattered by $\mathcal{H}$ if for any subset D' of D there is an $h \in \mathcal{H}$ such that $h(x) = 1$ for all $x \in D'$ and $h(x) = 0$ for all $x \in D - D'$. That is, patterns in $\mathcal{H}$ can be used to define all possible bi-partitions of D. For many interesting pattern spaces, the VC-dimension scales polynomially with the data dimension d. The following result exploits this property by showing that if a space has VC-dimension that is polynomial in d, then the space is sample-efficient learnable in the PAC-RPAD model.

Theorem 2. *For any pattern space $\mathcal{H}$ with finite VC-dimension $\mathcal{V}_\mathcal{H}$,* RAREPATTERNDETECT *is PAC-RPAD with sample complexity* $N(\delta, \epsilon) = O\left(\frac{1}{\epsilon^2}\left(\mathcal{V}_\mathcal{H} \log \frac{1}{\epsilon^2} + \log \frac{1}{\delta}\right)\right)$.

Proof. When the pattern space $\mathcal{H}$ is infinite, we want to bound the probability $P\left(\sup_{h\in\mathcal{H}} |\widehat{f}(h) - f(h)| > \frac{\epsilon}{2}\right)$, which can be achieved by bounding $P\left(\sup_{h\in\mathcal{H}} |\widehat{P}(h) - P(h)| > \frac{\epsilon}{2} U_{min}\right)$, where, $\widehat{P}(h)$ is an estimate of $P(h)$ based on sampled data. Let, $\epsilon_f = \frac{\epsilon}{2} U_{min}$. Using the VC uniform convergence bound on frequency estimates [6, Thm. 12.5] we have

$$P\left(\sup_{h\in\mathcal{H}} |\widehat{P}(h) - P(h)| > \epsilon_f\right) \leq 8\mathcal{S}_\mathcal{H}(N) e^{-N\epsilon_f^2/32}. \quad (1)$$

where, $\widehat{P}(h)$ is an estimate based on N i.i.d. samples from $\mathcal{P}$ and $\mathcal{S}_\mathcal{H}(N)$ is the Shatter Coefficient, which is the largest number of subsets that can be formed by intersecting some set of N points with patterns from $\mathcal{H}$.

Now, for any $N > 2\mathcal{V}_\mathcal{H}$, we can bound the Shatter Coefficient as: $\mathcal{S}_\mathcal{H}(N) < (\frac{eN}{\mathcal{V}_\mathcal{H}})^{\mathcal{V}_\mathcal{H}}$ [6, Thm. 13.3]. Hence, from Equation 1 we have

$$P\left(\sup_{h\in\mathcal{H}} |\widehat{P}(h) - P(h)| > \epsilon_f\right) < 8(\frac{eN}{\mathcal{V}_\mathcal{H}})^{\mathcal{V}_\mathcal{H}} e^{-N\epsilon_f^2/32}. \quad (2)$$

We want to bound this probability by δ, which yields

$$N \geq \frac{32}{\epsilon_f^2}\left(\mathcal{V}_\mathcal{H} \log(N) + \mathcal{V}_\mathcal{H} \log \frac{e}{\mathcal{V}_\mathcal{H}} + \log \frac{8}{\delta}\right). \quad (3)$$

Using the fact that $\log(N) \leq \alpha N - \log(\alpha) - 1$, where, $N, \alpha > 0$ and setting $\alpha = \frac{\epsilon_f^2}{64\mathcal{V}_\mathcal{H}}$, we get

$$\begin{aligned}\frac{32\mathcal{V}_\mathcal{H}}{\epsilon_f^2} \log(N) &\leq \frac{32\mathcal{V}_\mathcal{H}}{\epsilon_f^2}\left(\frac{\epsilon_f^2}{64\mathcal{V}_\mathcal{H}} N - \log \frac{\epsilon_f^2}{64\mathcal{V}_\mathcal{H}} - 1\right) \\ &= \frac{N}{2} + \frac{32\mathcal{V}_\mathcal{H}}{\epsilon_f^2} \log \frac{64\mathcal{V}_\mathcal{H}}{\epsilon_f^2 e}. \quad (4)\end{aligned}$$

Applying results from Equation 4 into Equation 3 and substituting the original value of ϵ_f we prove the Theorem 2:

$$N \geq \frac{256}{\epsilon^2} \frac{1}{U_{min}^2}\left(\mathcal{V}_\mathcal{H} \log\left(\frac{256}{\epsilon^2} \frac{1}{U_{min}^2}\right) + \log \frac{8}{\delta}\right).$$

□

5 APPLICATION TO SPECIFIC PATTERN SPACES

Most state-of-the-art anomaly detectors assign an anomaly score to data points and then detect points based on a score threshold or present a ranked list to the user. Further, while not usually explained explicitly, the scores are often based on frequency estimates of patterns in some space $\mathcal{H}$ with rare patterns leading to higher anomaly scores. While RAREPATTERNDETECT was designed as a pure implementation of this principle, it is reasonable to expect that its sample complexity is related qualitatively to the sample complexity of other algorithms grounded in pattern frequency estimation.

In this section, we consider a number of pattern spaces underlying existing algorithms and show that the sample complexity of those spaces is small. The spaces are thus all learnable in the PAC-RPAD framework, which offers some insight into why existing algorithms often show strong performance even in high-dimensional spaces.

5.1 CONJUNCTIONS

Consider a space $\mathcal{X}$ over d Boolean attributes and a pattern space $\mathcal{H}$ corresponding to conjunctions of those attributes. This setup is common in the data mining literature, where each boolean attribute corresponds to an "item" and a conjunction corresponds to an "item set", which indicates which items are in the set. Rare pattern mining has been studied for such spaces and applied to anomaly detection [1, 16]. In this case, the pattern space has a finite size 2^d and thus by Theorem 1 is efficiently PAC-RPAD learnable with sample complexity $O\left(\frac{1}{\epsilon^2}\left(d + \log \frac{1}{\delta}\right)\right)$. If we limit attention to conjunctions of at most k attributes, then the sample complexity drops to $O\left(\frac{1}{\epsilon^2}\left(k \log(d) + \log \frac{1}{\delta}\right)\right)$, which is sub-linear in d.

5.2 HALFSPACES

Given a continuous space $\mathcal{X} \subseteq \Re^d$, a *half space pattern* is an oriented d-dimensional hyperplane. A data point satisfies a half space pattern if it is on the positive side of the hyperplane. The half-space mass algorithm [5] for anomaly detection operates in this pattern space. Roughly speaking, the algorithm assigns a score to a point x based on the frequency estimates of random half spaces that contain x. The VC-dimension of d-dimensional half spaces is well known

to be $d+1$ and hence this space is sample-efficient learnable in the PAC-RPAD model.

5.3 AXIS ALIGNED HYPER RECTANGLES

For a continuous space $\mathcal{X} \subseteq \Re^d$, an axis-aligned hyper rectangle (bounded or unbounded) can be specified as a conjunction of threshold tests on a subset of the dimensions. The pattern space of axis-aligned rectangles are often implicit in decision tree algorithms, where each internal tree node specifies one threshold test.

The state-of-the-art anomaly detection algorithms, Isolation Forest [10] and RS-Forest [18] (among others), are based on this space. The core idea of these algorithms is to build a forest of T random decision trees, where each node specifies a random threshold test. The trees are grown until either a maximum depth is reached or a leaf node contains only a single data point. Each leaf corresponds to an axis-aligned hyper rectangle. Given a point x, the algorithm can compute the leaf node it reaches in each tree, yielding a set of T hyper-rectangle patterns. The score for x is then based on the average score assigned to each pattern, which is related to the pattern frequency, dimension, and volume.

The VC-dimension of the space of axis parallel hyper rectangles in $\Re^d$ is $2d$ [3]. Thus, the pattern space underlying Isolation Forest and RS-Forest is sample-efficient learnable in the PAC-RPAD model.

5.4 STRIPES

A stripe pattern in $\Re^d$ can be thought of as an intersection of two parallel halfspaces with opposite orientations and can be defined by the inequalities: $a \leq w^\top x \leq a + \Delta$, where, $w \in \Re^d$, a and $\Delta \in \Re$ and Δ represents the width of the stripe. The stripe pattern space consists of the set of all such stripes.

The very simple, but effective, anomaly detector, LODA [11], is based on the stripes pattern space. The main idea of LODA is to form a set of T sparse random projections along some random directions in the subspaces of $\Re^d$ and then estimate a discretized 1D histogram based on the projected values. Each bin of each histogram can be viewed as corresponding to a stripe in the original $\Re^d$ space with orientation defined by the direction of the random projection and location/width defined by the bin localtion/width.

To the best of our knowledge the VC-dimension of the stripes pattern space has not been previously derived. A loose bound can be found by noting that stripes are a special case of the more general pattern space of intersections of half spaces and then applying the general result for bounding the VC-dimension of intersections [3], which gives an upper bound on the VC-dimension of stripes of $4\log(6)(d+1) = O(d)$. Hence, stripes are PAC learnable in the RPAD model.

5.5 ELLIPSOIDS AND SHELLS

Anomaly detectors based on estimating "local densities" often form estimates based on frequencies observed in ellipsoids around query points. In particular, an ellipsoid pattern in a d dimensional space can be represented by $(x-\mu)^\top A\ (x-\mu) \leq t$, where $t \in \Re$, $\mu \in \Re^d$ and A is a $d \times d$ positive definite symmetric matrix. An upper bound for the VC-dimension of ellipsoids in $\Re^d$ is $(d^2+3d)/2$ [2]. Hence the ellipsoid pattern space is sample-efficient learnable in the PAC-RPAD model. However, we see that the complexity is quadratic in d rather than linear as we saw above for spaces based on hyperplane separators.

A related pattern space is the space of *ellipsoidal shells*, which is the analog of stripes for ellipsoidal patterns. An ellipsoidal shell in $\Re^d$ can be thought of as the subtraction of two ellipsoids with the same center and shape, but different volumes. That is, the shell is a region defined by $t \leq (x-\mu)^\top A(x-\mu) \leq t+\Delta$, where $\Delta \in \Re$ is the width. Shells naturally arise as the discretized level sets of multi-dimensional Gaussian density functions, which are perhaps the most widely-used densities in anomaly detection. We are unaware of previous results for the VC-dimensions of shells and show below that it is also $O(d^2)$.

Theorem 3. *The VC-dimension of the ellipsoidal shell pattern space in $\Re^d$ is upper bounded by* $2\log(6)(d^2+3d+2)$.

Proof. We can represent an ellipsoidal shell in $\Re^d$ as:

$$t \leq (x-\mu)^\top A\ (x-\mu)\ \leq t+\Delta. \tag{5}$$

Rewriting the equation of an ellipsoid in $\Re^d$:

$$\begin{aligned} & (x-\mu)^\top A\ (x-\mu)\ \leq t \\ \implies & x^\top A\, x - 2x^\top A\, \mu \leq t - \mu^\top A\, \mu \\ \implies & \sum_{i,j=1}^{d} A_{ij}x_i x_j - \sum_{i=1}^{d} 2(A\mu)_i x_i \leq t - \mu^\top A\mu \\ \implies & w^\top z \leq b. \end{aligned} \tag{6}$$

where, $w, z \in \Re^{d(d+1)/2+d}$. The vector w is a new parameter vector constructed from the original parameters. The matrix A gives $d(d+1)/2$ unique parameters, since A is symmetric, and the vector $A\mu$ gives another d parameters. The vector z is a new input constructed from the original input x, and $b = t - \mu^\top A\mu$.

Now, Applying result of Equation 6 to Equation 5 we get

$$\begin{aligned} & t - \mu^\top A\, \mu \leq w^\top z \leq t + \Delta - \mu^\top A\, \mu \\ \implies & t' \leq w^\top z \leq t' + \Delta. \end{aligned} \tag{7}$$

The Equation 7 is a representation of a stripe in $\Re^{d(d+1)/2+d}$. So, we apply the same approach from previous Section 5.4 i.e. the case of two halfspaces, which gives the upper bound of $4\log(6)(d(d+1)/2+d+1) = 2\log(6)(d^2+3d+2) = O(d^2)$. □

6 EXPERIMENTS

The above results suggest one reason for why state-of-the-art anomaly detectors often perform significantly better than random, even on high-dimensional data. However, existing empirical work does not go much further in terms of providing an understanding of learning curves for anomaly detection. To the best of our knowledge, there has not been a significant study of how the empirical performance of anomaly detectors varies as the amount of training/reference data increases. Typically empirical performance is reported for benchmark data sets without systematic variation of training set size, though other factors such as anomaly percentage are often varied. This is in contrast to empirical studies in supervised learning, where learning curves are regularly published, compared, and analyzed.

In this section, we present an initial investigation into empirical learning curves for anomaly detection. We are interested in the following experimental questions: 1) Will the empirical learning curves demonstrate the fast convergence predicted by the PAC-RPAD framework, and how is the convergence impacted by the data dimension and pattern space complexity? 2) How does the RPAD approach compare to a state-of-the-art detector based on the same pattern space on anomaly detection benchmarks? 3) In what ways do empirical learning curves for anomaly detection differ qualitatively from learning curves for supervised learning? While a complete empirical investigation is beyond the scope of this paper, below we provide experiments that shed light on each of these questions.

6.1 PATTERN SPACE AND ANOMALY DETECTOR SELECTION

A recent large scale evaluation [7] has shown that the Isolation Forest (IF) is among the top performing anomaly detectors across a wide range of benchmarks. This motivates us to focus our investigation on IF's pattern space of axis aligned hyper rectangles (see above). In order to allow for the complexity of this pattern space to be varied, we define *REC(k)* to be the space of all axis aligned hyper rectangles defined by at most k threshold tests on feature values.

The first step of IF is to construct a random forest of trees that are limited to a user-specified maximum leaf depth of k. Each tree leaf in the forest corresponds to a single pattern in REC(k). Thus, the first step of IF can be viewed as generating a random pattern space $\mathcal{H}_k \subseteq \text{REC}(k)$ that contains all leaf patterns in the forest. IF then operates by using training data to compute empirical frequencies $\widehat{P}(h)$ of patterns in $\mathcal{H}_k$ and then for any test point x computes an anomaly score based on those frequencies as follows (smaller is more anomalous):

$$\text{IF}(x) = \sum_{h \in \mathcal{H}_k[x]} d_h + c(\widehat{P}(h))$$

where, $d_h \leq k$ is the number of tests in pattern h and $c(h)$ is a function of the empirical frequency of h.[1]

In order to directly compare IF to our RPAD approach we will conduct multiple experiments, each one using a different randomly generated pattern space $\mathcal{H}_k$. We can then compute the scoring function corresponding to RAREPATTERNDETECT on those pattern spaces and compare to the performance of the IF scoring function. In particular, RAREPATTERNDETECT effectively assigns a score to x based on the minimum frequency pattern as follows:

$$\text{MIN}(x) = \min_{h \in \mathcal{H}_k[x]} \widehat{f}(h)$$

where the normalized frequency estimate $\widehat{f}(h)$ is normalized by a uniform density $U(h)$ over a region of bounded support defined by the training data. This normalizer is proportional to the volume of h and is easily computed. We see that compared to the IF scoring function with sums/averages over functions of each pattern in $\mathcal{H}_k[x]$, the MIN scoring function is only sensitive to the minimum frequency pattern. In order to provide a baseline in between these two, we also compare to the following alternative scoring function that averages normalized frequencies. This scoring function is given by

$$\text{AVE}(x) = \frac{1}{|\mathcal{H}_k[x]|} \sum_{h \in \mathcal{H}_k[x]} \widehat{f}(h).$$

AVE is included in order to observe whether averaging is a more robust approach to using normalized frequencies compared to MIN.

6.2 LEARNING CURVE GENERATION

We generate learning curves using three existing anomaly detection benchmarks:

Covertype [18]: $d = 10$ features, approximately 286k instances 0.9% anomalies.

Particle [7]: $d = 50$ features, approximately 130k instances with 5% anomalies.

Shuttle [7]: $d = 9$ features, approximately 58k instances with 5% anomalies.

These datasets were originally derived from UCI supervised learning benchmarks [9] by treating one or more classes as the anomaly classes and sub-sampling at an appropriate rate to produce benchmarks with certain percentages of anomalies. We have conducted experiments on a number of additional benchmarks, which are not included

[1] In particular, $c(h)$ is an estimate of the expected number of random tests required to completely isolate training data points that satisfy h, which is a function of the number of training points that satisfy h [10]. This score is motivated by attempting to estimate the "isolation depth" of x, which is the expected length of a random path required to isolate a point. Intuitively smaller isolation depths indicate a more anomalous point since it is easier to separate from other points.

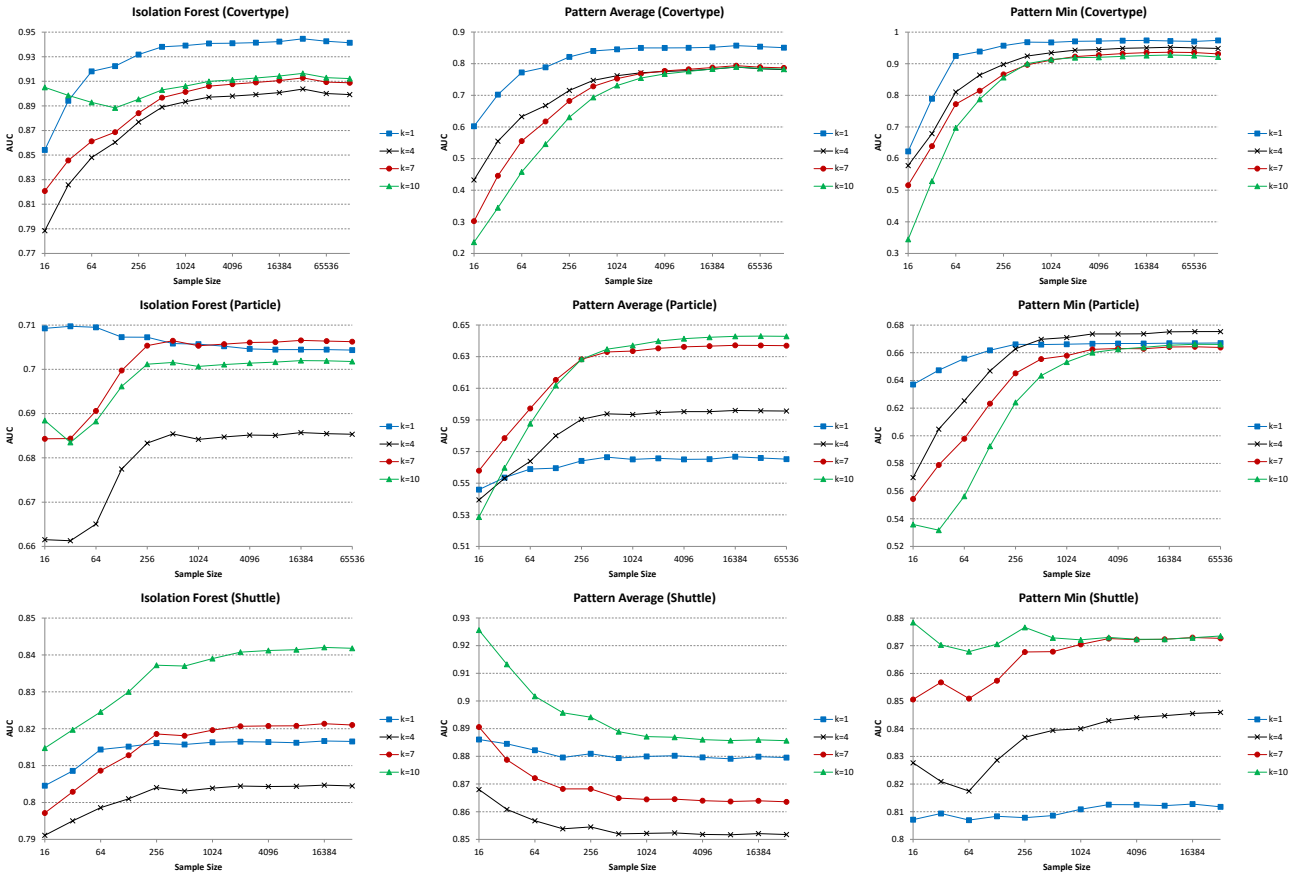

Figure 1: Learning Curves for the Three Scoring Methods (IF, AVE, MIN) with Varying Pattern Space Complexity k Over Three Benchmarks (Rows). MIN Represents the Main RPAD Approach Analyzed in this Paper.

for space reasons. These three data sets were selected as being representative of the qualitative learning curve types that we have observed. The data sets are divided into three sets of data: pattern generation data, training data, and test data for which we ensure that the fraction of anomaly points in each data set is approximately the same as for the full benchmark.

Given a benchmark and a specified pattern complexity k, we generate learning curves for each algorithm as follows. First, we use IF to generate a random pattern space $\mathcal{H}_k$, based on the pattern generation data using a forest of 250 random trees. Next for each desired training set size, we sample a training set of the appropriate size and use that data to estimate frequencies over $\mathcal{H}_k$, which can be done efficiently by passing each data point through each tree. Next, the scores defined above for IF, MIN, and AVE are computed for each test instance based on the frequency estimates and the *Area Under the Curve (AUC)* of those scores is computed relative to the ground truth anomalies. This process is repeated 50 times for each training set size and the resulting AUCs are averaged to produce a final learning curve.

6.3 RESULTS

Figure 1 gives learning curves for each benchmark (rows) and each of the anomaly detectors IF, MIN, and AVE (columns). Recall that MIN represents the main RPAD approach analyzed in this paper. In each case, four learning curves are shown for pattern space complexities $k = 1, 4, 7, 10$ and training instances range from 16 to 2^{15}.

Convergence rate: Each learning curve for each algorithm and benchmark follows a trajectory starting at 16 training instances to a converged performance at 2^{15} points. We see that in all cases, convergence to the final performance occurs very quickly, in particular around 1024 samples. This convergence rate is not significantly impacted by the dimensionality of the data, which can be seen by comparing the results for Particle with $d = 50$ to the other data sets with $d = 9$ and $d = 10$. Rather we see that the convergence rate visibly depends on the pattern space complexity k, though to a relatively small degree. In particular, we see that the convergence for the simplest space $k = 1$ tends to be faster than for $k = 10$ across our experiments. These observations agree with our analysis. The VC-dimension of REC(k), which controls worst case convergence, is dominated by the limiting effect of k. These observations are

consistent across additional benchmarks not shown here.

Relative Algorithm Performance: Here we focus on comparing the different detectors, or scoring functions, in terms of their converged performance. For the Cover and Shuttle data sets we see that the converged performance of MIN is better or competitive than the converged performance of IF for all values of k. For the Particle data set, the IF scoring function outperforms MIN consistently by a small margin. This shows that despite its simplicity the RPAD approach followed by MIN appears to be competitive with a state-of-the-art detector based on the same pattern space. Experiments on other benchmarks, not shown, further confirm this observation.

The converged performance of AVE tends to be worse than both IF and MIN for Covertype and Particle and is slightly better on Shuttle. It appears that for these data sets (and others not shown) that averaging is not significantly more robust than minimizing and can even hurt performance. One reason for degraded performance is that AVE can be influenced by the cumulative effect of a large number of non-rare patterns, which may sometimes overwhelm the signal provided by rare patterns.

An interesting observation is that for each data set, the best performing pattern space (i.e. value of k) is usually the same across the different learning algorithms. For example, for Covertype, $k = 1$ yields the best converged performance for all scoring functions. This observation, which we also frequently observed in other data sets, suggests that the choice of pattern space can have a performance impact that is as large or larger than the impact of the specific scoring function used. To understand this, note that the performance of an anomaly detector depends on both the convergence of its scoring function and the match between the scoring function and the semantic notion of anomaly for the application. The pattern space choice has a large influence on the this match since it controls the fundamental distinctions that can be made among data points.

Qualitative Properties: The qualitative behavior of the learning curves exhibits a couple of nonintuitive properties compared to supervised learning curves. First, in supervised learning, we typically expect and observe that more complex hypothesis spaces converge to a performance that is at least as good as simpler spaces, though more complex spaces may underperform at small sample sizes due to variance. This does not appear to be the case for anomaly detection learning curves in general. For example, the performance of the simplest pattern space ($k = 1$) on Covertype converges to better performance than the more complex spaces. This has also been observed in other data sets and does not appear to be due to premature termination of the learning curve. Rather, we hypothesize that this behavior is due to the mismatch between the anomaly detection scores and the semantic notion of anomaly in the benchmark. In particular, rare patterns in REC(1) are apparently a better indicator of the semantic notion of anomaly than some distracter rare patterns in REC(10). Indeed, it is straightforward to construct synthetic examples with such behavior.

Another nonintuitive aspect is that, in at least two cases, the learning curves consistently decrease in performance, while typically in supervised learning, ideal learning curves are non-decreasing. The most striking example is the performance of AVE on Shuttle, where all learning curves steadily decrease. After further analysis, this type of behavior again appears to be explained by the mismatch between the semantic notion of anomalies and the scoring function. In particular, the variance across different trials of the learning curve for large sample sizes is much smaller than for small sample sizes. It also turns out that the distribution of scoring functions generated for the small sample sizes is skewed toward solutions that better match the ground truth anomalies compared to the converged scoring function. Thus, the average performance for small sample sizes is better. We have observed this decreasing learning curve behavior in other data sets as well, though it is much more common for learning curves to increase.

7 SUMMARY

This work is motivated by the observation that many statistical anomaly detection methods perform much better than random in high dimensions using relatively small amounts of training data. Our PAC-RPAD framework attempts to explain these observations by quantifying the sample complexity in terms of the complexity of the pattern spaces underlying such anomaly detectors. Our results mirror those in supervised learning by showing that the VC-dimension of the pattern spaces is the dominating factor that controls sample complexity. We showed that for several state-of-the-art detectors, the underlying pattern spaces had well-behaved VC-dimensions with respect to the data dimensionality, which offers a partial explanation for their good performance. On the empirical side, we investigated for the first time, to the best of our knowledge, learning curves for anomaly detection. The experiments confirmed the fast convergence predicted by the theory. The results also suggest that our simple algorithm, which was shown to be PAC-RPAD, is competitive with the state-of-the-art algorithm Isolation Forest when using the same pattern space. Finally, the learning curves showed some interesting qualitative differences compared to supervised learning curves. In particular, the results highlight the importance of selecting a pattern space that is likely to be a good match to the semantic notion required for an application.

Acknowledgements
This work is partially supported by the Future of Life Institute and DARPA under contract number FA8650-15-C-7557 and W911NF-11-C-0088.

References

[1] M. Adda, L. Wu, and Y. Feng. Rare itemset mining. In *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*, pages 73–80. IEEE, 2007.

[2] A. Auger and B. Doerr. *Theory of randomized search heuristics: Foundations and recent developments*, volume 1. World Scientific, 2011.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[4] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[5] B. Chen, K. M. Ting, T. Washio, and G. Haffari. Half-space mass: a maximally robust and efficient data depth method. *Machine Learning*, 100(2-3):677–699, 2015.

[6] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[7] A. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. 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, 2013.

[8] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[9] M. Lichman. UCI machine learning repository, 2013.

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

[11] T. Pevnỳ. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, 2016.

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

[13] C. D. Scott and R. D. Nowak. Learning minimum volume sets. *The Journal of Machine Learning Research*, 7:665–704, 2006.

[14] C. Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, In press.

[15] M. A. Siddiqui, A. Fern, T. G. Dietterich, and W.-K. Wong. Sequential feature explanations for anomaly detection. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*. ACM, 2015.

[16] L. Szathmary, A. Napoli, and P. Valtchev. Towards rare itemset mining. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, volume 1, pages 305–312. IEEE, 2007.

[17] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[18] K. Wu, K. Zhang, W. Fan, A. Edwards, and P. S. Yu. RS-Forest: a rapid density estimator for streaming anomaly detection. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 600–609. IEEE, 2014.