Unsupervised Learning of Distributional Properties can Supplement Human Labeling and Increase Active Learning Efficiency in Anomaly Detection

Jaturong Kongmanee [1] **Mark Chignell** [1] **Khilan Jerath** [2] **Abhay Raman** [2]

Abstract

Exfiltration of data via email is a serious cybersecurity threat for many organizations. Detecting data exfiltration (anomaly) patterns typically requires labeling, most often done by a human annotator, to reduce the high number of false alarms. Active Learning (AL) is a promising approach for labeling data efficiently, but it needs to choose an efficient order in which cases are to be labeled, and there are uncertainties as to what scoring procedure should be used to prioritize cases for labeling, especially when detecting rare cases of interest is crucial. We propose an adaptive AL sampling strategy that leverages the underlying prior data distribution, as well as model uncertainty, to produce batches of cases to be labeled that contain instances of rare anomalies. We show that (1) the classifier benefits from a batch of representative and informative instances of both normal and anomalous examples, (2) unsupervised anomaly detection plays a useful role in building the classifier in the early stages of training when relatively little labeling has been done thus far. Our approach to AL for anomaly detection outperformed existing AL approaches on three highly unbalanced UCI benchmarks and on one real-world redacted email data set.

1. Introduction

Data exfiltration is an unauthorized process of transferring an individual's or organization's sensitive data outside an organization's perimeter. Exfiltrating data via email is an often-used method and is a serious cybersecurity threat for many organizations, irrespective of whether carried out by organized crime, commercial competitors, external bad actors, or careless or malicious insiders. Sensitive data can be transmitted as plain text in an email body, or attached as a file. There are several solutions available for combating data exfiltration, but they come with their own shortcomings. For example, if email domains are blacklisted, a determined insider could easily circumvent this by setting up accounts with different domains. Securing email gateways (SEGs) may be effective in blocking phishing emails; however, they can't stop all spear phishing emails, targeted phishing attacks using social engineering to impersonate trustworthy insiders to trick them into revealing login credentials, installing malware, or stealing data. Rule-Based solutions using "if-then" statements and regular expressions to look for data exfiltration signals are impossible to maintain because patterns and sensitivity in data change over time.

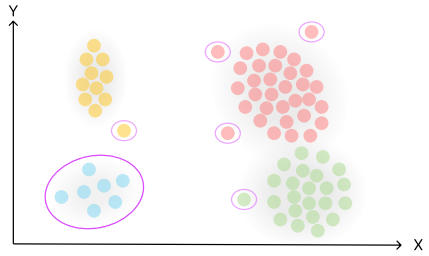


Figure 1. A simple example of anomalies (highlighted with ellipses) in a two-dimensional data set. The large ellipse shows a well-defined group of anomalies, while the ellipses around single points highlight anomalies lying in low-density regions.

In cases where defense at the perimeter is insufficient (as is the case if malicious acts are able to compromise accounts) methods are needed to identify activities such as email data exfiltration (the anomalous instance) that may be carried out using compromised accounts. Anomalous patterns are dynamic in nature and the current notion of normal patterns might not be representative in the future (Chandola et al., 2009; Hodge & Austin, 2004). Thus, defining a precise decision boundary between normal and anomalous patterns is extremely difficult and is domain-specific. In practice, it is not known if dense regions consist of only normal examples and anomalous examples are those residing in low-density regions near the decision boundary. Moreover, it is also possible that anomalous examples potentially reside

[1]University of Toronto [2]Sun Life Financial. Correspondence to: Jaturong Kongmanee <jaturong.kongmanee@mail.utoronto.ca>.

ICML workshop on AI & HCI, the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA, 2023.

in clusters occupying low-density regions (e.g., the circled cluster of points shown in the lower left of Figure 1).

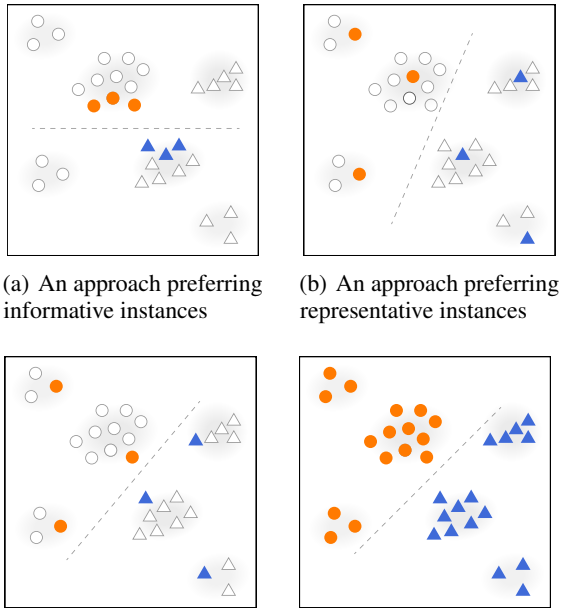
Detecting anomalies typically requires labeling, most often done by a human annotator, to build a classifier that can capture evolving anomalous patterns. However, the labeling process tends to be expensive both in terms of time and cost, and active learning (AL) methods are used to more efficiently take human knowledge into account. AL is a branch of machine learning where the key idea is to sample a small proportion of the data and obtain labels for that sample from a human annotator.

In practice, fully labeled data may not be required since good model performance is often obtained when models are trained on a well-selected subset of the data. Considerable research has demonstrated that AL can produce more efficient labeling of subsets. For instance, (Settles, 2011; 2009) used AL to maintain model performance while reducing the size of the labeled training set. Early AL methods have generally assumed that prior class probabilities are similar (balanced classes), which is not realistic in anomaly detection where the proportion of anomalies is extremely low. Sampling strategies based on model uncertainty are widely used but result in an over-reliance on cases near the decision boundary, where there is a danger that the human judge may be no more confident about her labels than the model is about its predictions.

AL is also influenced by the cold start problem (Houlsby et al., 2014; He & Garcia, 2009; Konyushkova et al., 2017; Gao et al., 2020) which potentially limits performance for uncertainty-based sampling when the initial training set is limited. One issue of particular concern is the likelihood of sampling biases influencing the model when there are few labels to guide it, where the model may ignore some regions of the sample spaces or even completely overlook certain classes.

Unsupervised ML anomaly detection techniques do not suffer from the cold start problem because they leverage the underlying data distribution rather than labels. Unsupervised ML anomaly detection techniques can improve the detection of new patterns of anomalous and rare examples, but labeling (explicit supervision) is still required in most cases to reduce the high number of false alarms that might otherwise occur.

In this paper, we ask the question, can we combine unsupervised and supervised methods to increase the efficiency of AL and to reduce the amount of human labeling effort required while still achieving a reasonable level of anomaly detection performance? Our answer to this question focuses on enhancing the sampling strategy for AL. We show empirically that the enhanced sampling method outperforms three baseline methods in terms of the area under the precision-recall curve (PRAUC) and anomaly detection rate on each of four data sets (three highly unbalanced UCI data sets and one redacted email provided by a financial company).



(a) An approach preferring informative instances

(b) An approach preferring representative instances

(c) Balancing informativeness and representativeness

(d) Optimal decision boundary trained on the entire data set

Figure 2. A conceptual illustration of sampling instances of linearly separable data. The white circles and triangles represent unlabeled samples. The orange circles and the blue triangles are labeled as positive and negative, respectively.

2. Related Work

Several efforts have been made to improve sampling strategies in AL. Sampling strategies can be measured based on informativeness, representativeness, or a combination of both. Informativeness refers to the extent to which querying a sample can reduce the model uncertainty. In contrast, representativeness measures how well a sample represents the underlying distribution of unlabeled data (Settles, 2009). Sampling the most informative instances has been used extensively, with strategies including query-by-committee (Dagan & Engelson, 1995; Freund et al., 1997; Seung et al., 1992), uncertainty sampling (Balcan et al., 2007; Lewis & Catlett, 1994; Lewis, 1995; Tong & Koller, 2001), and optimal experimental design (Flaherty et al., 2005; Yu et al., 2006). The main disadvantage of these strategies is that they ignore the prior data distribution, which can be useful for AL. The selection of query samples in the initial rounds of AL is based on only a few labeled examples, and can lead to sample bias if the distributional properties of the data are ignored, as shown in Figure 2a. This problem is especially noticeable when dealing with data that is highly unbalanced.

When a class is rare, its representative cases in that class may be overlooked because the data distribution of that class cannot be estimated with the relatively few instances sampled. Consequently, potential anomaly examples residing in classes associated with low-density regions may be ignored.

When sampling using an unsupervised approach, representativeness measures utilize the cluster structure of unlabeled data and focus on selecting the most suitable instances to represent the unlabeled data (Nguyen & Smeulders, 2004; Dasgupta & Hsu, 2008). Locally linear reconstructions are used to identify the data samples that adequately reconstruct the entire data set (Zhang et al., 2011). Without utilizing classification uncertainty (since labels are not used), the effectiveness of this approach is highly dependent on the performance of clustering results. As shown in Figure 2b, the representative sampling selects instances lying at the centers of clusters and can approximate accurate decision boundaries, but many queries are required. In practice, unsupervised methods need to be supplemented with labeling (supervision) at some point so that the model can converge to a sufficient level of classification performance. Previous work reported by (Huang et al., 2010; Ebert et al., 2012; Kremer et al., 2014) shows that using only one sampling strategy for AL may lead to a reduction in performance.

Early AL algorithms tried to find the optimal query examples by combining informativeness and representativeness measures. In Figure 2c, balancing both strategies potentially yields a decision boundary close to the optimal (Figure 2d) with fewer labels. In (Xu et al., 2003), the authors proposed a sampling strategy that performs clustering on the instances that are near a decision boundary. One limitation of this approach is its inability to exploit unlabeled examples that are more distant from the decision boundary. (Thrun & Möller, 1991) used an approach that switched randomly between uncertainty sampling and random sampling. (Nguyen & Smeulders, 2004) dynamically balanced uncertainty and the density of instances using a sampling strategy that preclustered data with the k-medoids algorithm. However, the method developed in (Nguyen & Smeulders, 2004) does not account for unbalanced classes, and the density estimation for each data point is limited to only the current set of clusters. (Pelleg & Moore, 2004; Stokes et al., 2008) proposed using a fixed combination of low likelihood and high uncertainty criteria for anomaly detection.

Ideally, AL sampling methods should adapt to the amount of "knowledge" that an ML model has about the distribution of cases, and about the relationship between the type of label and the position of instances in the feature space. Non-adaptive sampling criteria (such as those mentioned in the preceding paragraph) do not adjust the scoring criteria as the number of labeling samples increases and learning progresses. For instance, the model may waste effort by sampling near the current decision boundary, where there is often a high degree of uncertainty in the labels and where labeling may add little additional value/information. While fully automated models have achieved some success, they lack flexibility in terms of possible time-varying trade-off between an unsupervised approach (useful when there are only a few labels) and a supervised approach (likely better when the model is better trained). Given that there is this trade-off, it would likely be useful to allow a human annotator to control the behavior of the sampling strategy, at least in some scenarios.

3. Method

Sampling based on informativeness measures selects cases residing in low-density regions near the decision boundary of the current model. This approach will be able to find outlier anomalies, but is not capable of finding anomalies that are located in clusters. Clustering the data can be helpful in two ways. First, the representative samples located at the center of clusters are more significant than others and should be prioritized for labeling. Second, samples within the same cluster are likely to have the same label (Blum & Chawla, 2001). Thus, sampling based on representativeness measures can identify sufficiently accurate decision boundaries, but many calls to query labels from a human annotator are required. Therefore, we design our sampling strategy to be adaptive with a time-varying trade-off. Initially, the strategy is biased towards unsupervised methods (e.g., all ten in a batch of 10 instances for labeling are selected using an unsupervised method). In successive sampling rounds, there is increasing use of supervised methods. In the formulation used here (see section 3.1.3), parameters determine how quickly the transition from predominantly unsupervised to predominantly supervised sampling occurs during successive rounds. By parameterizing the scoring criterion, the human annotator is given control of the parameter settings that specify the trade-off between the number of unsupervised and supervised instances in each AL round. In the remainder of this section, we first introduce the batch mode AL approach and then list the heuristics that guide a potentially more effective AL approach.

3.1. Active Learning

In this section, we formally describe the AL approach that will be used. Given a classifier $f(\mathbf{x};\theta)$, unlabeled samples $\mathcal{U}$, a labeled training set $\mathcal{L}$, and input $\mathbf{x} \in \mathcal{U}$, a sampling strategy $\phi(\mathbf{x}, f(\mathbf{x};\theta))$ is a function of $\mathbf{x}$ and $f(\cdot)$ that the AL uses to select samples for labeling:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in U} \phi(\mathbf{x}, f(\mathbf{x};\theta)), \tag{1}$$

A batch mode AL selects batches of b instances at a time for human annotation to obtain an accurate model at a lower

labeling cost than regular supervised learning. The standard AL procedure is as follows:

1. Select a set of unlabeled instances $\mathcal{L} \subset \mathcal{U}$ for labeling.
2. Train a classifier $f(\mathbf{x}; \theta)$ with $\mathcal{L}$
3. Select $\mathbf{x}^* \in \mathcal{U} \setminus \mathcal{L}$ for labels using $\phi(\cdot)$
4. Assign labels $\mathbf{y}^*$ to $\mathbf{x}^*$ and update pools of labeled and unlabeled samples
5. Repeat steps $2-5$ until the classifier's performance is achieved or a number of iterations have reached a predefined number.

3.1.1. Representative Sampling

Our approach to AL includes representative sampling, where the goal is to learn the underlying prior data distribution of the unlabeled data and to select batches of representative samples in the early stage of AL. Mixture density estimation is used to delineate important regions (including anomalous examples residing in clusters with low-density regions) of the sample space to avoid sampling biases that may occur where a small number of labeled instances are not representative of the overall data set. Thus, we propose to use unsupervised analysis of the multivariate data distribution to reduce bias in the early stages of training. This is done by representing density variations in the space that guide anomaly detection when labeled instances are rare.

In this work, we represent the underlying data distribution presumably generated by a mixture model in terms of the Gaussian Mixture Model (GMM). The GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities (Dempster et al., 1977; McLachlan & Basford, 1988). Compared to other clustering methods, such as K-Means, a GMM provides statistical inferences concerning the underlying distributions that can be used later to determine the degree of anomaly of samples (Aggarwal & Aggarwal, 2017; Wang et al., 2019; Yang et al., 2021). We expect different Gaussian components in the mixture to learn different distributions that correspond to a variety of data patterns. The parameters of the mixture model are estimated by maximum likelihood estimate (MLE) via the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Our method works as follows:

1. Identify the number of mixture components ($\mathcal{K}$) corresponding to which of the alternative formulations has the lowest Bayesian Information Criterion (BIC) score (Schwarz, 1978).
2. Fit GMM with $\mathcal{K}$ components using EM.
3. Return n_{repr} centroids with the lowest probability density if $\mathcal{K} \geq n_{repr}$, where n_{repr} is the number of instances being selected in a batch, or else return $n_{repr} + (n_{repr} - \mathcal{K})$, where $(n_{repr} - \mathcal{K})$ is the instances with the lowest likelihood presumably generated by the learned distribution.

Our representative sampling returns centroids, including potential anomaly centroids residing in clusters with low-density regions that can approximate the underlying data distribution. Ranking each sample in order of increasing model likelihood and selecting the most anomalous instances to minimize model variances, can improve refinement of decision boundaries.

3.1.2. Informative Sampling

Our sampling approach uses information entropy (Shannon, 1948) as a measure of uncertainty/informativeness. This method selects the most informative samples, i.e., the samples that are close to the decision boundary, presumably passing through low-density regions of the marginal data distribution. The least informative samples are those where one of the classes has a high probability (examples far away from the decision boundary). Formally, for k-class classification, the information entropy $\mathbb{H}(x)$ of sample x can be defined as $\mathbb{H}(x) = -\sum_{i=1}^{k} P(y_i|x) \cdot logP(y_i|x)$, where $P(y_i|x)$ is the probability that the current sample x is predicted to be class y_i. The greater the entropy of the sample, the greater its uncertainty, which we refer to as *Max Entropy*. However, a batch mode AL strategy that selects multiple informative samples each time might result in samples that are very similar, providing little information. Thus, the selected batch should be informative for the model, while being diverse enough to minimize redundancy between sampled instances. Our method operates in the following way:

1. Select the top $\frac{n_{info} \times 100}{b}\%$ most informative instances from the pool of unlabeled data $\mathcal{U}$, where n_{info} is the number of instances being selected in a batch.
2. Apply K-Means clustering to all informative instances obtained from the previous step to identify n_{info} groups. The k-means++ seeding algorithm (Arthur & Vassilvitskii, 2007) is used to promote diversity among these informative instances.
3. Return n_{info} instances that are closest to the cluster centroids for human labeling.

In summary, our proposed informative sampling heuristic avoids the selection of redundant instances and concentrates on the most important informative instances in selecting samples.

3.1.3. ADAPTIVE SAMPLING

Our approach aims to combine representative and informative samplings as a function of AL iterations. The proposed approach prioritizes representative sampling in the initial phase and linearly[1] balances both criteria until informative sampling becomes dominant. This ensures there is always a mixture of the two criteria in the early AL stage. Since human experience is a valuable resource and should be incorporated into solving a problem, we allow a human annotator to control the behavior of sampling strategies to improve the model performance. The *balancing function* has the form:

$$\alpha(t, b, c, T_1, T_2) = \begin{cases} b, 0 & t < T_1 \\ \underbrace{b - \mathcal{B}(\cdot)}_{n_{repr}}, \underbrace{\mathcal{B}(\cdot)}_{n_{info}} & T_1 \leq t < T_2; t = t - T_1 \\ 0, b & T_2 \leq t \end{cases} \tag{2}$$

where t is the AL iteration, b is the batch size, $c \in [0, 1]$ is a human annotator's confidence level for her initial classifier, T_1 is the iteration to start balancing (i.e., adding in informative samples of some cases), T_2 is the stopping iteration (i.e., using only the informative sampling), $\mathcal{B}(\cdot)$ is $\mathbf{mod}(t + \lceil b * c \rceil, b)$, and the function returns two values: n_{repr} and n_{info}. n_{repr} is the number of instances selected through the representative sampling, while n_{info} is based on informative sampling, where the sum of these values equals b. Pseudo-code for the proposed algorithm is given in Alg. 1.

When there is access to a sufficiently large training set, a human annotator can modify parameter c accordingly. For instance, setting c to 0.5 results in a batch that consists of samples selected by both criteria (i.e., a $50/50$ supervised and unsupervised rounds) starting from the first iteration, as opposed to having the unsupervised approach dominating in the first iteration. We hypothesize that switching between supervised and unsupervised training based on the amount of labeled instances already available, and model uncertainty associated with the current level of training, will be useful in creating more efficient AL.

4. Experiments

Table 1. Dataset Statistics

DATA SET	DIMENSIONS	SAMPLES	ANOMALIES (%)
ABALONE	9	1920	29(1.50%)
ANN-THYROID-1V3	21	3251	73(2.25%)
CARDIOTOCOGRAPHY	22	1700	45(2.65%)
REDACTED EMAIL	42	672	418(62.20%)

[1]We also experimented with exponential and polynomial, but we found linearly transitioning between two criteria worked best consistently across all experiments.

Algorithm 1 Adaptive AL Sampling Strategy

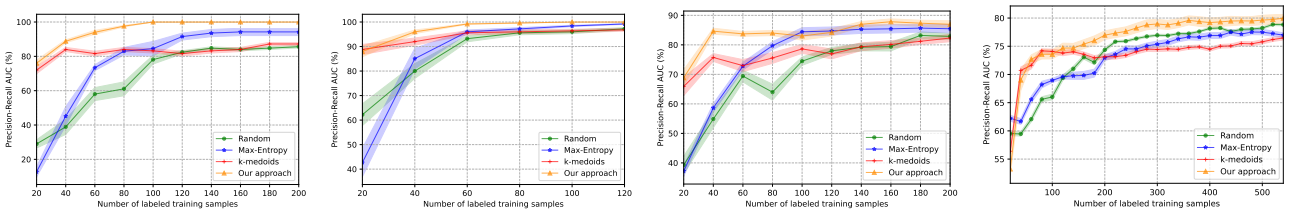
Require: unlabeled data set $\mathcal{U}$, labeled set $\mathcal{L}$, batch size b, initial number of labeled examples $\mathcal{M}$, number of iterations $\mathcal{T}$, classifier $f(\mathbf{x}; \theta)$, sampling strategy $\phi(\cdot)$, balancing function $\alpha(\cdot)$, iteration to start balancing T_1, iteration to stop balancing T_2, confidence level c.

Labeled data set $\mathcal{L} \leftarrow \mathcal{M}$ examples drawn uniformly at random from $\mathcal{U}$ along with queried labels.
Train an initial classifier θ_0 on $\mathcal{L}$
for $t = 1, 2, \ldots, \mathcal{T}$ **do**
  $n_{repr}, n_{info} \leftarrow \alpha(t, b, c, T_1, T_2)$ ▷ See 2
  $\mathcal{X}_{repr} \leftarrow \mathsf{Repr}(n_{repr}, \phi(\theta_{t-1}, \mathcal{U}))$ ▷ Section 3.1.1
  $\mathcal{X}_{info} \leftarrow \mathsf{Info}(n_{info}, \phi(\theta_{t-1}, \mathcal{U}))$ ▷ Section 3.1.2
  $\hat{\mathcal{X}} \leftarrow \{\mathcal{X}_{repr}, \mathcal{X}_{info}\}$
  Query labels for $\hat{\mathcal{X}}$
  $\mathcal{L} \leftarrow \mathcal{L} \cup \hat{\mathcal{X}}$
  $\mathcal{U} \leftarrow \mathcal{U} \setminus \hat{\mathcal{X}}$
  $\theta_t \leftarrow \text{Train}(\theta_{t-1}, \mathcal{L})$ ▷ Train a classifier on a newly updated training set.
end for
return The model $\theta_{\mathcal{T}}$

4.1. Experimental setup

Data sets. Following (Das et al., 2016; Zong et al., 2018), we evaluate our method on the following highly unbalanced UCI benchmarks (Asuncion & Newman, 2007) used for anomaly detection: *Abalone*, *Thyroid (ANN-Thyroid)*, *Cardiotocography*, and on one real-world *redacted email* data set (tabular data) provided by a financial service company. The *redacted email* data set had 42 features, including variables such as binary variables that indicate whether certain sensitive terms are present in the subject line or attachment names (full details provided in (Wang et al., 2023)). A number of anomalies and normal examples in each dataset are shown in Table 1. We note that while our method focuses on the highly unbalanced data sets, improving the sampling strategy, in general, will further improve AL for the case of the more balanced data set (i.e., redacted email), as observed in section 4.2 that our method benefits from carefully selecting important instances.

Baselines and method. We compared our method with the following baselines: **i) Random:** The naive baseline of selecting a batch of size b uniformly at random from the unlabeled pool at each round for labeling. This baseline allows us to compare the benefit of AL over passive learning. **ii) Max Entropy:** A widely used informative sampling strategy baseline that selects a batch of b informative instances according to the entropy of the example's predictive class probability distribution. For binary classification, max entropy is equivalent to margin sampling and least confident sampling approaches (Settles, 2009). **iii) k-medoids:** A



(a) Abalone (b) ANN-Thyroid (c) Cardiotocography (d) Redacted Email

Figure 3. PRAUC, on four different data sets, compared against size of training set (accumulating number of instances sampled).

robust-to-noise unsupervised anomaly detection technique that selects b medoids. In contrast to K-Means, k-medoids use data points in a data set as estimates of central location instead of centroids (means), which may not belong to the clusters. Also, k-medoids is less influenced by outliers and noise, making it more robust than K-Means. Previous research by (Syarif et al., 2012; Agrawal & Agrawal, 2015) shows that k-medoids produces better results than K-Means in detecting novel network anomalies in cybersecurity.

Evaluation metrics. We used a threshold-invariant metric, the area under the precision-recall curve (PRAUC), which is suitable for rare binary events and unaffected by model specificity (Davis & Goadrich, 2006), and has been shown to be more informative than AUROC score when the classes are highly unbalanced (Saito & Rehmsmeier, 2015). We also plotted a total number of true anomalies discovered as a function of number of queries presented to the human annotator. Ideally the number of true anomalies identified should increase quickly and is thus a measure of the quality of AL performance. Another reason for expecting number of true anomalies to increase quickly is that we want to make efficient use of the human annotator.

Implementation details. Unless otherwise specified, in all experiments, we use Support Vector Machines (SVMs) with a Radial Basis Function (RBF) kernel as a classifier due to its well-understood theoretically (Kremer et al., 2014). We calibrate the probabilities of a classifier using Platt scaling (Platt et al., 1999). We divided each data set into two sets using Stratified Shuffle Split [2] to preserve the same percentage for each class as in the original data set. All sampling strategies were performed on the unlabeled set (80%), and the effectiveness of the sampling strategies was evaluated after each batch based on the other unseen fixed set (20%) referred to as the test set. We considered a hard case of AL, where we started with two randomly selected labeled examples per class and set a confidence level $(c = 0)$. We set $T_1 = 0$, $T_2 = 5$, and evaluated them in a batch mode AL setup with a batch size of $b = 20$. The batch sample size of 20 had been found to be the maximum that could be implemented with human annotators in a previous survey study by (Wang et al., 2023) at a large financial services company. All sampling strategies started with the same initial labeled set, unlabeled set, and test set. The experiments were repeated for 50 independent runs, and mean performance, with 95% confidence intervals, are reported.

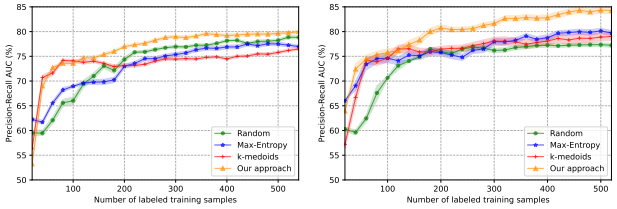
[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html

4.2. Experimental results and discussion

PRAUC performance on the cold start problem. In Figure 3, we demonstrate the empirical results with four initial labeled instances (two for each class) across all data sets and baselines. Our method outperforms Max Entropy sampling by focusing on learning different distributions that correspond to a variety of data patterns, without overlooking a potential rare class, to more effectively estimate decision boundaries within the early AL stage. It can be seen that higher performance was achieved across data sets using our method, presumably due to a reduction in sampling bias. Compared to k-medoids, our method starts with a competitive level of performance and converges more quickly to a high level of performance.

We hypothesize that our method benefits from the greater use of the proposed informative sampling in later AL rounds. Max Entropy outperforms k-medoids after a sufficient number of labeled samples are collected. However, unsupervised methods may not converge to sufficiently high levels of performance and even if they do, the labeling costs may be too high. So, our method provides a way to adjust the trade-off between unsupervised and supervised learning so that sampling bias can be reduced in earlier AL rounds (focusing on an unsupervised approach) while greater focus on labeled instances can efficiently enhance model performance in later rounds of AL. Our main contribution in this paper is that we provide a novel way to control the trade-off in AL between exploration of the feature space to avoid sampling bias (unsupervised learning) and learning from labeled instances (supervised learning).

Manually varying the trade-off between both sampling strategies. We further verify the flexibility and effectiveness

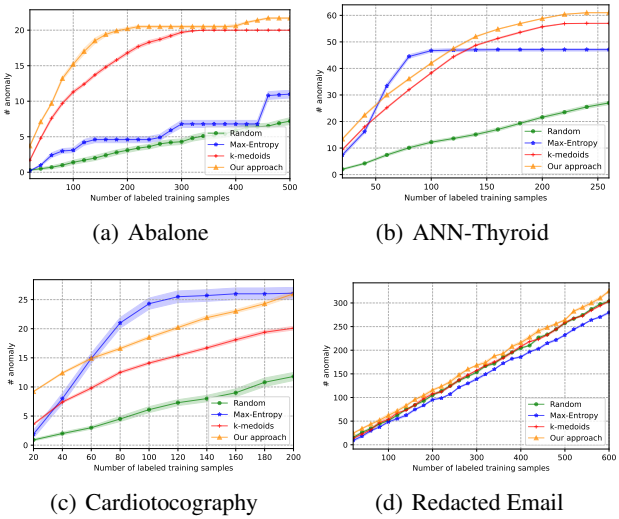


(a) $\mathcal{M} = 4$ (b) $\mathcal{M} = 20$

Figure 4. PRAUC on a redacted email test set, compared against the size of the training set for two settings: the amount of the initial labeled set ($\mathcal{M} = 4$) and ($\mathcal{M} = 20$).

of our method on a redacted email dataset. Figure 4a (identical to Figure 3d but repeated here for comparison purposes) shows that our method can mitigate the effects of sampling bias, as evidenced in Figure 3, by initially setting c to 0 (i.e., starting with the unsupervised learning that explores the data distribution). As expected, the PRAUC performance in the first iteration is lower than Max Entropy. However, the unsupervised technique used does not suffer from the cold start problem, and outperforms Max Entropy in later iterations. By the seventh iteration (where a total of 140 instances have been labeled), our method provides a batch that consists of samples purely selected by the proposed informative sampling, which leads to higher performance than k-medoids. Figure 4b shows the benefit of having access to a sufficiently large training set. In this setting, we adjust parameter c to 0.5 to obtain a batch of samples selected by both criteria in equal amounts, instead of having the unsupervised approach dominate from the beginning (i.e., $c = 0$). Our method closely matches the performance of Max Entropy in early rounds as the initial model has better knowledge about the feature space, demonstrating the benefit of incorporating human knowledge into controlling the behavior of sampling strategies. The proposed method at the seventh iteration selects non-redundant samples solely based on the informative measures as the amount of labeled examples increases, achieving higher performance than all baselines.

Anomaly detection rate. In this experiment, we compare how quickly algorithms can identify anomalous classes in a data set. This will help optimize the use of human annotators' time. The results are illustrated in Figure 5 for our method and the three existing approaches. Our method quickly identifies anomalous samples and is able to include true anomaly examples for human labeling from the first iteration, as opposed to Max Entropy and uniformly sampling approaches. All methods perform equally well for a redacted email data set. We hypothesize this is because the classes in this data set were balanced. Our method exhibits sample-efficient properties by demonstrating performance improvements (Figure 3b and 3c) while detecting fewer anomalies than Max Entropy in just a few iterations (Figure



(a) Abalone (b) ANN-Thyroid

(c) Cardiotocography (d) Redacted Email

Figure 5. Number of true anomalies discovered on four different data sets, compared against size of training set (accumulating number of instances sampled).

5b and 5c). We hypothesize that our method prevents the selection of redundant instances and instead focuses on the most important informative instances.

5. Conclusion

We have demonstrated that some of the widely used sampling strategies for AL perform poorly in practical scenarios where classes are unbalanced. Our proposed method works well in the presence of highly unbalanced classes and anomalies, as well as when anomalies are frequent. Our simulations show that the method proposed here leads to AL rounds where batches of samples contain instances of rare anomalies. Batches of instances that contain only one class (typically no anomalies when anomalies are rare) will not lead to much new information when cases are labeled. Thus in order to efficiently learn distinctions between anomalies and non-anomalies, there should be examples of anomalies in every batch, more effectively utilizing human annotator time in the labeling process. Our approach is aimed at increasing the sampling of rare classes, and it is flexible, since we do not assume a particular data distribution, making it applicable to a wide range of data sets. Our approach provides several indicators to assist a human annotator in identifying anomalous data, as well as controlling the behavior of sampling strategy in different settings.

Acknowledgements

We would like to thank Thanyathorn Thanapatheerakul and Worrawat Engchuan for helpful discussions.

References

Aggarwal, C. C. and Aggarwal, C. C. *An introduction to outlier analysis*. Springer, 2017.

Agrawal, S. and Agrawal, J. Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60:708–713, 2015.

Arthur, D. and Vassilvitskii, S. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.

Asuncion, A. and Newman, D. Uci machine learning repository, 2007.

Balcan, M.-F., Broder, A., and Zhang, T. Margin based active learning. In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13-15, 2007. Proceedings 20*, pp. 35–50. Springer, 2007.

Blum, A. and Chawla, S. Learning from labeled and unlabeled data using graph mincuts. 2001.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3): 1–58, 2009.

Dagan, I. and Engelson, S. P. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pp. 150–157. Elsevier, 1995.

Das, S., Wong, W.-K., Dietterich, T., Fern, A., and Emmott, A. Incorporating expert feedback into active anomaly discovery. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 853–858. IEEE, 2016.

Dasgupta, S. and Hsu, D. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 208–215, 2008.

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Ebert, S., Fritz, M., and Schiele, B. Ralf: A reinforced active learning formulation for object class recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3626–3633. IEEE, 2012.

Flaherty, P., Arkin, A., and Jordan, M. Robust design of biological experiments. *Advances in neural information processing systems*, 18, 2005.

Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133, 1997.

Gao, M., Zhang, Z., Yu, G., Arık, S. Ö., Davis, L. S., and Pfister, T. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 510–526. Springer, 2020.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Hodge, V. and Austin, J. A survey of outlier detection methodologies. *Artificial intelligence review*, 22:85–126, 2004.

Houlsby, N., Hernández-Lobato, J. M., and Ghahramani, Z. Cold-start active learning with robust ordinal matrix factorization. In *International conference on machine learning*, pp. 766–774. PMLR, 2014.

Huang, S.-J., Jin, R., and Zhou, Z.-H. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23, 2010.

Konyushkova, K., Sznitman, R., and Fua, P. Learning active learning from data. *Advances in neural information processing systems*, 30, 2017.

Kremer, J., Steenstrup Pedersen, K., and Igel, C. Active learning with support vector machines. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(4):313–326, 2014.

Lewis, D. D. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pp. 13–19. ACM New York, NY, USA, 1995.

Lewis, D. D. and Catlett, J. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pp. 148–156. Elsevier, 1994.

McLachlan, G. J. and Basford, K. E. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

Nguyen, H. T. and Smeulders, A. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 79, 2004.

Pelleg, D. and Moore, A. Active learning for anomaly and rare-category detection. *Advances in neural information processing systems*, 17, 2004.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10 (3):e0118432, 2015.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, pp. 461–464, 1978.

Settles, B. Active learning literature survey. 2009.

Settles, B. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pp. 1–18. JMLR Workshop and Conference Proceedings, 2011.

Seung, H. S., Opper, M., and Sompolinsky, H. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Stokes, J. W., Platt, J., Kravis, J., and Shilman, M. Aladin: Active learning of anomalies to detect intrusions. 2008.

Syarif, I., Prugel-Bennett, A., and Wills, G. Data mining approaches for network intrusion detection: from dimensionality reduction to misuse and anomaly detection. *Journal of Information Technology Review*, 3(2): 70–83, 2012.

Thrun, S. B. and Möller, K. Active exploration in dynamic environments. *Advances in neural information processing systems*, 4, 1991.

Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Wang, H., Bah, M. J., and Hammad, M. Progress in outlier detection techniques: A survey. *Ieee Access*, 7:107964–108000, 2019.

Wang, L., Giang, C., Jerath, K., Raman, A., Lie, D., Chignell, M., et al. Implementing active learning in cybersecurity: Detecting anomalies in redacted emails. *arXiv preprint arXiv:2303.00870*, 2023.

Xu, Z., Yu, K., Tresp, V., Xu, X., and Wang, J. Representative sampling for text classification using support vector machines. In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14–16, 2003. Proceedings 25*, pp. 393–407. Springer, 2003.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Yu, K., Bi, J., and Tresp, V. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pp. 1081–1088, 2006.

Zhang, L., Chen, C., Bu, J., Cai, D., He, X., and Huang, T. S. Active learning based on locally linear reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2026–2038, 2011.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.