1-2008

Rare Class Discovery Based on Active Learning

Jingrui He
Carnegie Mellon University

Jaime G. Carbonell
Carnegie Mellon University, jgc@cs.cmu.edu

Published In
.

Rare Class Discovery Based on Active Learning

Jingrui He and **Jaime Carbonell**
School of Computer Science
Carnegie Mellon University

Abstract

In machine learning, the new-class discovery problem remains an open challenge, especially for emergent rare classes. However, the challenge is of crucial importance for applications such as detecting new financial fraud patterns, new viral mutations and new network malware, most of which 'hide' among vast volumes of normal data and observations. This paper focuses on a new approach, based on local-topology density estimation, applicable to discovering examples of the rare classes rapidly, despite non-separability with the majority class(es). The new method, called ALICE, and its variant MALICE, are shown effective both theoretically and empirically in outperforming other methods in the literature, both on challenging synthetic data and on real data sets.

1 Introduction

Supervised machine learning methods require labeled training examples for each class (Mitchell 1997). Semi-supervised methods such as co-training (Blum & Mitchell 1998), and active learning (Donmez & Carbonell 2007) share the same requirement, although the former also utilizes unlabeled examples, and the latter optimizes sampling strategies to obtain additional labels. However, both assume that at least one or more instance of each class is given – i.e., they do not address the new-class discovery challenge.

In many real world problems, we are interested in rapidly discovering examples of rare classes, which are known to be existent in the data set a priori. Often times very small rare classes obfuscated to appear as members of known majority classes. For instance, the vast majority of financial transactions are legitimate, but a small number may be fraudulent; detecting early instances of new fraud patterns is a major first step towards systematically finding and stopping such illicit activity (Bay *et al.* 2006). Another example is network intrusion detection; systematically finding the early onset of new malicious network activities among huge volumes of routine network traffic is a critical unmet challenge (Wu *et al.* 2007). If we sample the data at random, we will need to examine a very large number of routine majority-class examples before discovering the emergent rare classes. This problem is also a bottleneck in reducing the sample complexity of active learning (Balcan, Beygelzimer, & Langford 2006) (Dasgupta 2005).

Compared with the rich literature on unbalanced-category classification, up until now, only a few methods have been proposed to address the rare class discovery challenge. For example, in (Pelleg & Moore 2004), the authors assumed a mixture model to fit the data, and selected examples for labeling according to different criteria; in (Fine & Mansour 2006), the authors proposed a generic consistency algorithm, and proved upper bounds and lower bounds for this algorithm in some specific situations. Scalability in new-class discovery was addressed in (Carbonell *et al.* 2006). Online new-topic assignment for documents in a stream was proposed in (Blei & Lafferty 2006). Whereas the above evidence a recent surge of interest in new class discovery, these methods in general require that the majority classes and the rare classes be separable or nearly-separable to work well. However, in real applications, the support regions of the majority classes and the rare classes often overlap strongly (sometimes due to intentional obfuscation).

In this paper, we propose a new active learning method for rapid rare-class discovery, named ALICE. It works in the cases where we know the existence of some rare classes, but do not have any labeled examples from these classes. Different from existing methods on class exploration, in our method, the rare classes may overlap with the majority classes. However, different rare classes should be distinct from each other. Intuitively, for each class, ALICE makes use of the local topology defined by nearest neighbors to measure local density around each example based on class-specific radii. Then it selects an example with the maximum change in local density on a certain scale, asks an external oracle for its label, and gradually increases the scale until it finds

an example from that class. The core ALICE is proven to be effective theoretically. In practice, to avoid repeatedly sampling the same class once discovered, we have modified ALICE to produce MALICE, which performs much better than existing methods in our set of experiments, both on synthetic and real data.

The rest of the paper is organized as follows. In Section 2, we introduce ALICE and MALICE with theoretical justifications. Then we give some experimental results to demonstrate their effectiveness in Section 3. Finally, we conclude the paper in Section 4.

2 Class Exploration Method

2.1 Notation

Given a set of unlabeled examples $S = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$, which come from m distinct classes, i.e. $y_i \in \{1, \ldots, m\}$. For the sake of simplicity, assume that there is one majority class with prior p_1, which corresponds to $y_i = 1$, and all the other classes are rare classes with priors $p_2, \ldots, p_m$, $p_1 \gg p_i$, $i \neq 1$. Without loss of generality, suppose that we are only interested in the rare classes, and the goal is to find at least one example from each rare class by requesting as few total labels as possible.

2.2 Method

The proposed method ALICE is presented in Algorithm 1. ALICE works as follows: Given the priors for the rare classes, we first estimate the number K_i of instances from class i in the set S. Then, for class i, at each example, we record its distance from the K_i^{th} nearest neighbor, which could be realized efficiently by kd-trees (Moore 1991) for medium or low input-space dimensionality. The minimum distance over all the examples is the class specific radius, and is assigned to r_i'. Next, we draw a hyper-ball centered at example x_j with radius r_i', and count the number of examples enclosed by this hyper-ball, which is denoted as n_j^i. n_j^i is roughly in proportion to the local density. To find examples from class i, in each iteration of Step 10, we subtract the local density of neighboring points from n_j^i, and let the maximum value be the score of x_j. The example with the maximum score is selected for labeling by the oracle. If the example is from class i, stop the iteration; otherwise, enlarge the neighborhood where the scores of the examples are re-calculated and continue.

To intuitively understand ALICE, assume that the rare classes are concentrated in small regions and the probability density function (pdf) of the majority class is locally smooth. Firstly, since the support regions of the rare classes are very small, it is important to find their scales. The r_i' values obtained in Step 3 will be used to calculate the local density n_j^i. Since r_i' is based on the minimum K_i^{th} nearest neighbor distance, it is never too large to smooth out changes of local density, and thus it is a good measure of the scale to begin with. Secondly, in each iteration of Step 8, the score of a certain point, corresponding to the change in local density, is the maximum of the difference in local density between this point and all of its neighboring points. In this way, we are not only able to select points on the boundary of the rare class i, but also points in the interior, given that the support region of class i is small. Finally, by gradually enlarging the neighborhood where the scores are calculated, we can further explore the interior of the support region, and increase our chance of finding rare class examples.

Algorithm 1 Active Learning for Initial Class Exploration (ALICE)

Require: S, $p_2, \ldots, p_m$

1: Initialize all the rare classes as undiscovered.
2: **for** $i = 2 : m$ **do**
3: Let $K_i = np_i$, where n is the number of examples.
4: For each example, calculate the distance between this example and its K_i^{th} nearest neighbor. Set r_i' to be the minimum value among all the examples.
5: **end for**
6: **for** $i = 2 : m$ **do**
7: $\forall x_j \in S$, let $NN(x_j, r_i') = \{x | x \in S, \|x - x_j\| \leq r_i'\}$, and $n_j^i = |NN(x_j, r_i')|$.
8: **end for**
9: **for** $i = 2 : m$ **do**
10: If class i has been discovered, continue.
11: **for** $t = 2 : n$ **do**
12: For each x_j that has been selected, $s_j^i = -\infty$; for all the other examples, $s_j^i = \max_{x_k \in NN(x_j, tr_i')} (n_j^i - n_k^i)$.
13: Select and query the label of $x = \arg\max_{x_j \in S} s_j^i$.
14: If the label of x is equal to i, break; otherwise, mark the class that x belongs to as discovered.
15: **end for**
16: **end for**

2.3 Justification

In this subsection, we prove that if the rare classes are concentrated in small regions and the pdf of the majority class is locally smooth, ALICE will repeatedly sample in the regions where rare class examples occur with high probability.

Let $f_i(x)$ denote the pdf of class i, where $i = 1, \ldots, m$ and $x \in \mathbb{R}^d$. To be precise, we make the following assumptions.

Assumptions

1. The pdf $f_i(x)$ of rare class i is uniform within a hyper-ball B_i of radius r_i [1] centered at b_i, $i = 2, \ldots, m$, i.e. $f_i(x) = \frac{1}{V(r_i)}$, if $x \in B_i$; and 0 otherwise, where $V(r_i) \propto r_i^d$ is the volume of B_i.
2. $f_1(x)$ is bounded and positive in B_i, $i = 2, \ldots, m$, i.e. $f_1(x) \geq \frac{c_{i1} p_i}{p_1 V(r_i)}$, $\forall x \in B_i$ and $f_1(x) \leq \frac{c_{i2} p_i}{p_1 V(r_i)}$, $\forall x \in \mathbb{R}^d$, where $c_{i1}, c_{i2} > 0$ are two constants.[2]

Furthermore, for each rare class i, $i = 2, \ldots, m$, let $r_{i2} = \frac{r_i}{(1+c_{i2})^{\frac{1}{d}}}$; and let $OV(\frac{r_{i2}}{2}, r_i)$ be the volume of the overlapping region of two hyper-balls: one is of radius r_i; the other one is of radius $\frac{r_{i2}}{2}$, and its center is on the sphere of the previous one. We have the following theorem, which proves the effectiveness of ALICE.

Theorem. If

1. For rare class i, $i = 2, \ldots, m$, let B_i^2 be the hyper-ball centered at b_i with radius $2r_i$. The minimum distance between the points inside B_i and the ones outside B_i^2 is not too large, i.e. $\max_{i=2}^m \min\{\|x_j - x_k\| | x_j, x_k \in S, \|x_j - b_i\| \leq r_i, \|x_j - b_i\| > 2r_i\} \leq \alpha$.
2. The rare classes are far apart, i.e. if $x_j, x_k \in S$, $\|x_j - b_i\| \leq r_i$, $\|x_k - b_{i'}\| \leq r_{i'}$, $i, i' = 2, \ldots, m$, and $i \neq i'$, then $\|x_j - x_k\| > \alpha$.
3. $f_1(x)$ is locally smooth, i.e. $\forall x, y \in \mathbb{R}^d, |f_1(x) - f_1(y)| \leq \frac{\beta \|x-y\|}{\alpha}$, where $\beta \leq \min_{i=2}^m \frac{p_i^2 OV(\frac{r_{i2}}{2}, r_i)}{2^{d+1} V(r_i)^2}$.
4. The number of examples is sufficiently large, i.e. $n \geq \max\{\max_{i=2}^m \frac{1}{2c_{i1}^2 p_i^2} \log \frac{3m-3}{\delta}$, $\max_{i=2}^m \frac{1}{2(1-2^{-d})^2 p_i^2} \log \frac{3m-3}{\delta}$, $\max_{i=2}^m \frac{1}{p_1^4 \beta^4 V(\frac{r_{i2}}{2})^4} \log \frac{3m-3}{\delta}\}$.

then with probability at least $1-\delta$, in every iteration of Step 8, after $\lceil \frac{2\alpha}{r_{i2}} \rceil$ rounds of Step 10, ALICE will query at least one example whose probability of coming from a rare class is at least $\frac{1}{3}$.

Proof. To prove the theorem, we need the following simple lemma.

Lemma. For each rare class i, $i = 2, \ldots, m$, $\forall \epsilon_i, \delta_i > 0$, if $n \geq \max\{\max_{i=2}^m \frac{1}{2c_{i1}^2 p_i^2} \log \frac{3m-3}{\delta}, \max_{i=2}^m \frac{1}{2(1-2^{-d})^2 p_i^2} \log \frac{3m-3}{\delta}, \max_{i=2}^m \frac{1}{\epsilon^4 V(\frac{r_{i2}}{2})^4} \log \frac{3m-3}{\delta}\}$, then with probability at least $1-\delta$, $\frac{r_{i2}}{2} \leq r_i' \leq r_i$ and $|\frac{n_j^i}{n} - E(\frac{n_j^i}{n})| \leq \epsilon V(r_i')$, $1 \leq j \leq n$.

Proof. First, notice that for each rare class i, the expected proportion of points falling inside B_i, $E(\frac{|NN(b_i, r_i)|}{n}) \geq (c_{i1}+1)p_i$, and that the maximum expected proportion of points falling inside any hyper-ball of radius $\frac{r_{i2}}{2}$, $\max_{x \in \mathbb{R}^d} [E(\frac{|NN(x, \frac{r_{i2}}{2})|}{n})] \leq 2^{-d} p_i$. Then

$$
\begin{aligned}
&\Pr[\exists i, \text{ s.t. } r_i' > r_i \text{ OR } \exists i, \text{ s.t. } r_i' < \frac{r_{i2}}{2} \\
&\text{OR } \exists i, \exists x_j \in S \text{ s.t. } |\frac{n_j^i}{n} - E(\frac{n_j^i}{n})| > \epsilon V(r_i')] \\
&\leq \sum_{i=2}^m \Pr[r_i' > r_i] + \sum_{i=2}^m \Pr[r_i' < \frac{r_{i2}}{2}] + \\
&\sum_{i=2}^m \Pr[r_i' \geq \frac{r_{i2}}{2} \text{ AND } \exists x_j \text{ s.t. } |\frac{n_j^i}{n} - E(\frac{n_j^i}{n})| > \epsilon V(r_i')] \\
&\leq \sum_{i=2}^m \Pr[|NN(b_i, r_i)| < K_i] \\
&+ \sum_{i=2}^m \Pr[\max_{x \in \mathbb{R}^d} |NN(x, \frac{r_{i2}}{2})| > K_i] \\
&+ \sum_{i=2}^m n \Pr[|\frac{n_j^i}{n} - E(\frac{n_j^i}{n})| > \epsilon V(r_i') | r_i' \geq \frac{r_{i2}}{2}] \\
&= \sum_{i=2}^m \Pr[|\frac{NN(b_i, r_i)}{n}| < p_i] \\
&+ \sum_{i=2}^m \Pr[\max_{x \in \mathbb{R}^d} |\frac{NN(x, \frac{r_{i2}}{2})}{n}| > p_i] \\
&+ n \sum_{i=2}^m \Pr[|\frac{n_j^i}{n} - E(\frac{n_j^i}{n})| > \epsilon V(r_i') | r_i' \geq \frac{r_{i2}}{2}] \\
&\leq \sum_{i=2}^m e^{-2nc_{i1}^2 p_i^2} + \sum_{i=2}^m e^{-2n(1-2^{-d})^2 p_i^2} \\
&+ 2n \sum_{i=2}^m e^{-2n\epsilon^2 V(r_i')^2}
\end{aligned}
$$

where the last inequality is based on Hoeffding bound.

Let $e^{-2nc_{i1}^2 p_i^2} \leq \frac{\delta}{3m-3}$, $e^{-2n(1-2^{-d})^2 p_i^2} \leq \frac{\delta}{3m-3}$ and $2ne^{-2n\epsilon^2 V(r_i')} \leq 2ne^{-2n\epsilon^2 V(\frac{r_{i2}}{2})^2} \leq \frac{\delta}{3m-3}$, we obtain $n \geq \frac{1}{2c_{i1}^2 p_i^2} \log \frac{3m-3}{\delta}$, $n \geq \frac{1}{2(1-2^{-d})^2 p_i^2} \log \frac{3m-3}{\delta}$, and $n \geq \frac{1}{\epsilon^4 V(\frac{r_{i2}}{2})^4} \log \frac{3m-3}{\delta}$. ■

Based on this lemma, using condition 4, let $\epsilon = p_1 \beta$, if the number of examples is sufficiently large, then with probability at least $1-\delta$, for each rare class i, $i =$

[1] This is the actual radius, as opposed to the class specific radius r_i'.

[2] Notice that here we are only dealing with the hard case where $f_1(x)$ is positive within B_i. In the separable case where the support regions of the majority class and the rare classes do not overlap, we can use other methods to detect the rare classes, such as the one proposed in (Pelleg & Moore 2004).

$2, \ldots, m$, $\frac{r_{i2}}{2} \leq r'_i \leq r$ and $|\frac{n^i_j}{n} - E(\frac{n^i_j}{n})| \leq p_1\beta V(r'_i)$, $1 \leq j \leq n$.

To better prove the theorem, given a point $x_j \in S$, we say that x_j is 'far from all the rare classes' iff for every rare class i, $\|x_j - b_i\| > 2r_i$, i.e. x_j is not within B^2_i. According to condition 3, $\forall x_j, x_k \in S$ s.t. x_j and x_k are far from all the rare classes and $\|x_j - x_k\| \leq \alpha$, $E(\frac{n^i_j}{n})$ and $E(\frac{n^i_k}{n})$ will not be affected by the rare classes. Therefore, in iteration i of Step 8 where we aim to find examples from rare class i, $|E(\frac{n^i_j}{n}) - E(\frac{n^i_k}{n})| \leq p_1\beta V(r'_i) \leq p_1\beta V(r_i)$. Furthermore, since α is always bigger than r_i, we have

$$\begin{aligned}
&|\frac{n^i_j}{n} - \frac{n^i_k}{n}| \\
&\leq |\frac{n^i_j}{n} - E(\frac{n^i_j}{n})| + |\frac{n^i_k}{n} - E(\frac{n^i_k}{n})| + |E(\frac{n^i_j}{n}) - E(\frac{n^i_k}{n})| \\
&\leq 3p_1\beta V(r_i) \qquad (1)
\end{aligned}$$

From inequality (1), it is not hard to see that $\forall x_j, x_k \in S$, s.t. x_j is far from all the rare classes and $\|x_j - x_k\| \leq \alpha$, $\frac{n^i_j}{n} - \frac{n^i_k}{n} \leq 3p_1\beta V(r_i)$, i.e. when $tr'_i = \alpha$,

$$\frac{s^i_j}{n} \leq 3p_1\beta V(r_i) \qquad (2)$$

This is because if x_k is not far from any of the rare classes, the rare classes may also contribute to $\frac{n^i_k}{n}$, and thus the score of x_j may be even smaller.

On the other hand, based on conditions 1 and 2, there exist two points $x_u, x_v \in S$, s.t. $\|x_u - b_i\| \leq r_i$, x_v is far from all the rare classes, and $\|x_u - x_v\| \leq \alpha$. Since the contribution of rare class i to $E(\frac{n^i_u}{n})$ is at least $\frac{p_i \cdot OV(\frac{r_{i2}}{2}, r_i)}{V(r_i)}$, so $E(\frac{x^i_u}{n}) - E(\frac{x^j_v}{n}) \geq \frac{p_i \cdot OV(\frac{r_{i2}}{2}, r_i)}{V(r_i)} - p_1\beta V(r'_i) \geq \frac{p_i \cdot OV(\frac{r_{i2}}{2}, r_i)}{V(r_i)} - p_1\beta V(r_i)$. Since for any example $x_j \in S$, we have $|\frac{n^i_j}{n} - E(\frac{n^i_j}{n})| \leq p_1\beta V(r'_i) \leq p_1\beta V(r_i)$, therefore

$$\begin{aligned}
\frac{n_u}{n} - \frac{n_v}{n} &\geq \frac{p_i \cdot OV(\frac{r_{i2}}{2}, r_i)}{V(r_i)} - 3p_1\beta V(r_i) \\
&\geq \frac{p_i \cdot OV(\frac{r_{i2}}{2}, r_i)}{V(r_i)} - \frac{3p_1 p_i^2 \cdot OV(\frac{r_{i2}}{2}, r_i)}{2^{d+1}V(r_i)}
\end{aligned}$$

Since p_i is very small, $p_i \gg \frac{6p_1p_i^2}{2^{d+1}}$; therefore, $\frac{n^i_u}{n} - \frac{n^i_v}{n} > \frac{3p_1p_i^2 \cdot OV(\frac{r_{i2}}{2}, r_i)}{2^{d+1}V(r_i)} \geq 3p_1\beta V(r_i)$, i.e. when $tr'_i = \alpha$,

$$\frac{s^i_u}{n} > 3p_1\beta V(r_i) \qquad (3)$$

In Step 10 of the proposed method, we gradually enlarge the neighborhood to calculate the change of local density to continue seeking an example of the rare class. When $tr'_i = \alpha$, based on inequalities (2) and (3), $\forall x_j \in S$ s.t. x_j is far from all the rare classes, we have $s^i_u > s^i_j$. Therefore, in this round of iteration, we will pick an example that is NOT far from one of the rare classes, i.e. there exists a rare class i_t s.t. the selected example is within $B^2_{i_t}$. Note that i_t is not necessarily equal to i, which is the rare class that we would like to discover in Step 8 of the method.

Finally, we show that the probability of picking an example that belongs to rare class i_t from $B^2_{i_t}$ is at least $\frac{1}{3}$. To this end, we need to calculate the maximum probability mass of the majority class within $B^2_{i_t}$. Consider the case where the maximum value of $f_1(x)$ occurs at b_{i_t}, and this pdf decreases by β every time x moves away from b_{i_t} in the direction of the radius by α, i.e. the shape of $f_1(x)$ is a cone in $(d+1)$ dimensional space. Since $f_1(x)$ must integrate to 1, i.e. $V(\frac{\alpha f_1(b_{i_t})}{\beta}) \cdot \frac{f_1(b_{i_t})}{d+1}$, where $V(\frac{\alpha f_1(b_{i_t})}{\beta})$ is the volume of a hyper-ball with radius $\frac{\alpha f_1(b_{i_t})}{\beta}$, we have $f_1(b_{i_t}) = (\frac{d+1}{V(\alpha)})^{\frac{1}{d+1}}\beta^{\frac{d}{d+1}}$. Therefore, the probability mass of the majority class within $B^2_{i_t}$ is:

$$\begin{aligned}
&V(2r_{i_t})(f_1(b_{i_t}) - \frac{2r_{i_t}}{\alpha}\beta) + \frac{2r_{i_t}}{\alpha}\frac{\beta}{d+1}V(2r_{i_t}) \\
&< V(2r_{i_t})f_1(b_{i_t}) = V(2r_{i_t})(\frac{d+1}{V(\alpha)})^{\frac{1}{d+1}}\beta^{\frac{d}{d+1}} \\
&= 2^d\frac{V(r_{i_t})}{(V(\alpha))^{\frac{1}{d+1}}}(d+1)^{\frac{1}{d+1}}\beta^{\frac{d}{d+1}} \\
&< (d+1)^{\frac{1}{d+1}}(2^{d+1}V(r_{i_t})\beta)^{\frac{d}{d+1}} \\
&\leq (d+1)^{\frac{1}{d+1}}(\frac{p_{i_t}^2 \cdot OV(\frac{r_{i_t2}}{2}, r_{i_t})}{V(r_{i_t})})^{\frac{d}{d+1}} < 2p_{i_t}
\end{aligned}$$

where $V(2r_{i_t})$ is the volume of a hyper-ball with radius $2r_{i_t}$. Therefore, if we select a point at random from $B^2_{i_t}$, the probability that this point is from rare class i_t is at least $\frac{p_{i_t}}{p_{i_t} + p_1 \cdot 2p_{i_t}} \geq \frac{p_{i_t}}{p_{i_t} + 2p_{i_t}} = \frac{1}{3}$. ■

2.4 Implementational Issues

According to our theorem, in each iteration of Step 8, with high probability, we may pick examples belonging to the rare classes after selecting a small number of examples. However, the discovered rare class i_t may not be the same as the rare class i that we hope to discover in this iteration of Step 8. Furthermore, we may repeatedly select examples from class i_t before finding one example from class i. To address these issues, we have modified the original ALICE algorithm to produce MALICE, which is shown in Algorithm 2.

There are two major differences between MALICE and ALICE. 1) In Step 12 of MALICE, once we have labeled an example, any unlabeled example within the class specific radius of this example will be precluded

Algorithm 2 Modified Active Learning for Initial Class Exploration (MALICE)

Require: $S, p_2, \ldots, p_m$

1: Initialize all the rare classes as undiscovered.
2: **for** $i = 2 : m$ **do**
3: Let $K_i = np_i$.
4: For each example, calculate the distance between this example and its K_i^{th} nearest neighbor. Set r'_i to be the minimum value among all the examples.
5: **end for**
6: Let $r'_1 = \max_{i=2}^{m} r'_i$.
7: **for** $i = 2 : m$ **do**
8: $\forall x_j \in S$, let $NN(x_j, r'_i) = \{x | x \in S, \|x - x_j\| \leq r'_i\}$, and $n_j^i = |NN(x_j, r'_i)|$.
9: **end for**
10: **for** $i = 2 : m$ **do**
11: If class i has been discovered, continue.
12: **for** $t = 2 : n$ **do**
13: For each x_j that has been selected, $\forall x_k \in S$, s.t. $\|x_j - x_k\| \leq r'_{y_j}$, $s_k^i = -\infty$; for all the other examples, $s_j^i = \max_{x_k \in NN(x_j, tr'_i)} (n_j^i - n_k^i)$.
14: Select and query the label of $x = \arg\max_{x_j \in S} s_j^i$.
15: If the label of x is equal to i, break; otherwise, $t = t - 1$, mark the class that x belongs to as discovered.
16: **end for**
17: **end for**

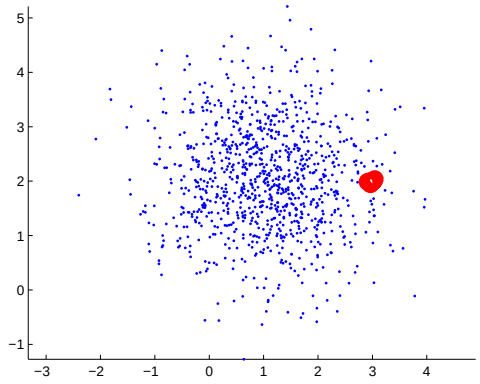
from selection. Since we have proved that with high probability, the class specific radius is less than the actual radius, this modification will help prevent examples of the same class from being selected repeatedly. 2) In Step 14 of MALICE, if the labeled example belongs to a rare class other than class i, we will not enlarge the neighborhood based on which the scores of the examples are re-calculated. This is to increase the chance that if tr'_i is close to α, we will select examples from B_i^2.

3 Experimental Results

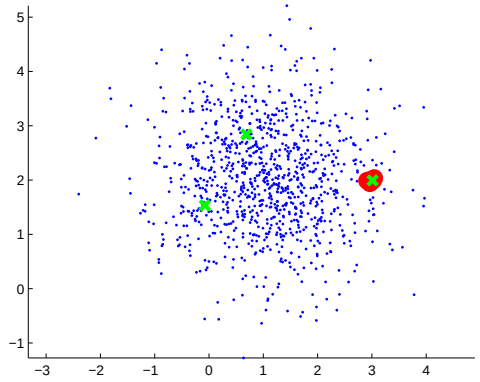
In this section, we compare MALICE with the best method proposed in (Pelleg & Moore 2004) (Interleave) and random sampling (RS) on both synthetic and real data sets. In Interleave, we use the number of classes as the number of components in the mixture model. For both Interleave and RS, we run the experiments 10 times and report the average results.

3.1 Synthetic data sets

Figure 1(a) shows a synthetic data set where there is only one rare class. The pdf of the majority class (shown in blue dots) is Gaussian and the pdf of the rare class (shown in red circles) is uniform within a small hyper-ball. There are 1000 examples from the majority class and only 10 examples from the rare class. Using Interleave, we need to label 35 examples on average; using RS, we need to label 101 examples on average; and using MALICE, we only need to label 3 examples in order to sample one example from the rare class we are interested in, which are denoted as 'x' in Figure 1(b). Notice that the first 2 examples that MALICE selects are not from the correct region. This is because the number of examples from the rare class is very small, and the local density may be affected by the randomness in the data.



(a) Data Set

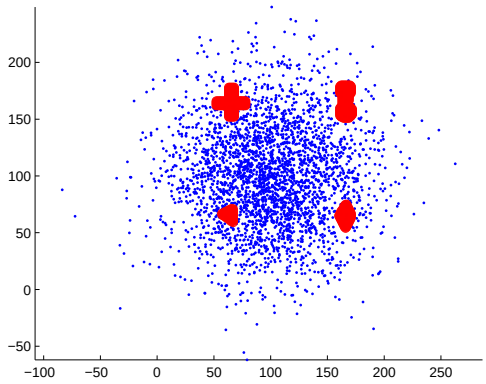


(b) Examples Selected by MALICE, denoted as 'x'

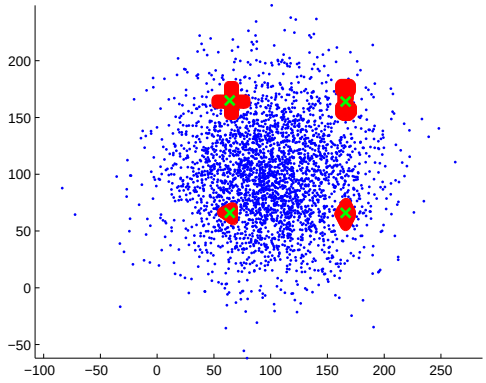
Figure 1: Synthetic Data Set 1.

In Figure 2(a), the majority class has 3000 examples (shown in blue dots) with Gaussian distribution. The 4 rare classes (shown in red circles) all have different shapes, and each has 267, 280, 84 and 150 examples respectively. Using Interleave, we need to label 382 examples on average; using RS, we need to label 68 examples on average; and using MALICE, we only need to label

4 examples, each of which is in a different rare class. The examples selected by MALICE are denoted as 'x' in Figure 2(b). Notice that with this dataset, Interleave is even worse than RS. This is because some rare class is within the dense region of the majority class. Therefore, it may take Interleave a long time to finally find one example from this rare class.



(a) Data Set



(b) Examples Selected by MALICE, denoted as 'x'

Figure 2: Synthetic Data Set 2.

3.2 Real data sets

In this subsection, we compare different methods on two real data sets: Shuttle (Brazdil & Gama 1991) and image data set. The first data set consists of 4515 examples, described by 9 dimensional features. The examples come from 7 classes: the proportion of the largest class (majority class) is 75.53%, and the proportion of the smallest class is 0.13%. The second data set consists of 5000 images, described by 244 dimensional features such as color and texture. The examples come from 6 classes: the proportion of the largest class (majority class) is 90.00%, and the proportion of the smallest class is 2.00%.

In Table 1 and 2, we compare the number of labeled examples for different methods on the two data sets respectively. From these tables, we can see that MALICE is significantly better than Interleave and RS: with Shuttle data set, to find all the rare classes, Interleave needs 132 label requests, RS needs 512 label requests, and MALICE only needs 84 label requests; with image data set, to find all the rare classes, Interleave needs 662 label requests, RS needs 112 label requests, and MALICE only needs 49 label requests. This is because as the number of components becomes larger, the mixture model generated by Interleave is less reliable due to the lack of labeled examples, thus we need to select more examples. Furthermore, the majority class and rare classes may not be nearly-separable, which is a disaster for Interleave. On the other hand, MALICE does not assume a generative model for the data, and only focuses on the change in local density, which is more effective on the two data sets.

Number of Rare Classes Discovered	1	2	3	4	5	6
MALICE	6	11	49	71	72	84
Interleave	1	52	107	109	115	132
RS	7	9	13	63	100	512

Table 1: The Number of Labeled Examples for Different Methods on Shuttle Data Set.

Number of Rare Classes Discovered	1	2	3	4	5
MALICE	3	4	14	43	49
Interleave	6	114	180	181	662
RS	10	22	39	61	112

Table 2: The Number of Labeled Examples for Different Methods on Image Data Set.

3.3 Conclusion

In this paper, we have proposed a new active learning method (ALICE) for rare-class discovery, which is a very important topic in many real problems, such as network intrusion detection and financial fraud detection. Different from existing methods, ALICE does not rely on the assumption that the data is nearly-separable. It works by selecting examples corresponding to regions with the maximum change in local density, and depending on scaling, it will select class-boundary or class-internal examples of the rare classes. ALICE could be scaled up using kd-trees (Moore 1991). The effectiveness of ALICE is guaranteed by theoretical justification, i.e. guarantees on the probability of discovering an example of rare classes, given a sampling strategy. Furthermore, to avoid repeatedly sampling in the same class in real applications, we have modified

ALICE accordingly to produce MALICE, which outperforms existing methods on both synthetic and real data sets. Future work involves studying the robustness of MALICE when the parameters provided to it (the number of rare classes, and the priors of each class) are unknown or just estimates.

References

Balcan, M.; Beygelzimer, A.; and Langford, J. 2006. Agnostic active learning. In *Proc. of the 23rd Int. Conf. on Machine Learning*, 65–72.

Bay, S.; Kumaraswamy, K.; Anderle, M.; Kumar, R.; and Steier, D. 2006. Large scale detection of irregularities in accounting data. In *Proc. of the 6th Int. Conf. on Data Mining*, 75–86.

Blei, D., and Lafferty, J. 2006. Dynamic topic models. In *Proc. of the 23rd Int. Conf. on Machine Learning*, 113–120.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proc. of the 23th Annual Conf. on Computational Learning Theory*, 92–100.

Brazdil, P., and Gama, J. 1991. Statlog repository. In *http://www.niaad.liacc.up.pt/old/statlog/datasets/shuttle/shuttle.doc.html*.

Carbonell, J.; Fink, E.; Jin, C.; Gazen, C.; Mathew, J.; Saxena, A.; Satish, V.; s. Ananthraman; Dietrich, D.; Mani, G.; Tittle, J.; and Durbin, P. 2006. Scalable data exploration and novelty detection. In *NIMD Workshop*.

Dasgupta, S. 2005. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 19*.

Donmez, P., and Carbonell, J. 2007. Dual-strategy active learning. In *Proc. of the 18th European Conf. on Machine Learning*.

Fine, S., and Mansour, Y. 2006. Active sampling for multiple output identification. In *The 19th Annual Conf. on Learning Theory*, 620–634.

Mitchell, T. 1997. *Machine Learning*. McGraw-Hill Science Engineering.

Moore, A. 1991. A tutorial on kd-trees. Technical report, University of Cambridge Computer Laboratory.

Pelleg, D., and Moore, A. 2004. Active learning for anomaly and rare-category detection. In *Advances in Neural Information Processing Systems 18*.

Wu, J.; Xiong, H.; Wu, P.; and Chen, J. 2007. Local decomposition for rare class analysis. In *Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 814–823.