Toward Interactive Search in Remote Sensing Imagery

Reid Porter[1], Don Hush, Neal Harvey, James Theiler
Los Alamos National Laboratory, Los Alamos, NM, USA.

ABSTRACT

To move from data to information in almost all science and defense applications requires a human-in-the-loop to validate information products, resolve inconsistencies, and account for incomplete and potentially deceptive sources of information. This is a key motivation for visual analytics which aims to develop techniques that complement and empower human users. By contrast, the vast majority of algorithms developed in machine learning aim to replace human users in data exploitation. In this paper we describe a recently introduced machine learning problem, called rare category detection, which may be a better match to visual analytic environments. We describe a new design criteria for this problem, and present comparisons to existing techniques with both synthetic and real-world datasets. We conclude by describing an application in broad-area search of remote sensing imagery.

Keywords: Rare category detection, interactive machine learning, anomaly detection, change detection.

1. INTRODUCTION

The typical problem in machine learning is to build a model that relates observed data X to a categorical variable $Y \in \{a, b, c, d, \dots\}$ that encodes higher level information. This problem is widely applicable: X could represent network data packets and Y could represent categories of malicious attacks, or X could represent persistent surveillance video and Y could represent categories of activities. Of particular interest, are applications where Y includes unknown or extremely rare categories. Sometimes, machine learning solutions to this problem involve an expert during the design phase to provide examples, or training data. But in nearly all cases, the user is not part of the final system. Of course in reality it is latter, in the exploitation environment, that users find themselves inevitably spending large amounts of time validating and correcting model outputs.

We propose a more direct approach that we call interactive machine learning. The long term objective is to develop machine learning methods that produce a final system that includes algorithms and users as components. This could produce algorithms that are optimal in terms of how they are used in practical exploitation environments, and could also provide these environments with a performance metric with which to optimize interfaces and visualization tools.

Rare category detection provides a step in this direction: instead of learning a model of the data, we learn a sampling strategy that determines which samples are shown to the user and with what probability. The learning objective is to minimize the number of samples required to observe at least one example from each category: $\{a, b, c, d, \dots\}$. Unlike the traditional machine learning objective which aims to identify all categories as accurately as possible, rare category detection produces models that bring new categories of information to the user's attention as quickly as possible. This approach assumes once a user sees a new category, the user will do the rest and either discount the information as uninteresting, or initiate follow-up analysis which would include developing more accurate models.

In practice, this approach is well matched to how many users are exploiting data. For example, in astronomical sky survey datasets the vast majority of the data (99.9%) is well explained by current theories and models. The remainder are anomalies, but 99% of these anomalies are uninteresting (due to sensor problems or artifacts) and only 1% of them (0.001% of the full dataset) are of scientific interest [1]. Traditional anomaly detection performs poorly since it is trying to identify all possible anomalies, and most of them are uninteresting. However in rare category detection, the system only needs to find one example of an interesting anomaly, since once identified, an expert will be more than willing to spend additional time to understand the new phenomena. We point out that astronomical data mining is also an application of interest to the visual analytics community [2].

[1] rporter@lanl.gov, 505-665-7508.

We begin in Section 2 with a brief discussion of an interactive machine learning system, and describe how this paper relates to a number of related works in rare category detection. In Section 3 we provide our main technical contribution, which is a new design criteria for rare category detection. We outline the key ideas and provide some results on benchmark datasets. In Section 4 we switch gears, and describe a visualization tool we have developed for broad-area search in remote sensing imagery. We discuss key features as well as future research that will be required to turn this interface into a practical interactive machine learning system.

2. BACKGROUND

In figure 1 we illustrate the building blocks of our long term objective: an interactive machine learning system. Data-driven and knowledge-driven modeling techniques translate raw data into a more abstract representation, it is general purpose and error prone. Users interact with the model to focus additional computation and provide context to resolve ambiguities and correct representation errors. The user interaction contains information about the user's preferences and priorities and also contains valuable domain knowledge about the dataset [3]. The interaction is recorded and translated into a user model using data-driven and knowledge-driven modeling techniques. Over time, and with sufficient end-user involvement, the user model provides a valuable tool for optimizing the data model, leading to increasingly specialized representations, and a narrowing of the semantic gap for particular users and applications.

In this paper we investigate a new design objective for the data-model called rare category detection. One of the more precise definitions of the rare category detection problem is given by [4] where it is called Multiple Output Identification. They provide an analysis of learning complexity for some instances of the problem. Authors in [5] propose a hierarchical mean shift procedure for model building and suggest a number of criteria for selecting samples. Our work has similar scope to [5] but focuses on a different criteria for model building, and to identify interesting samples, we simply use the model output.

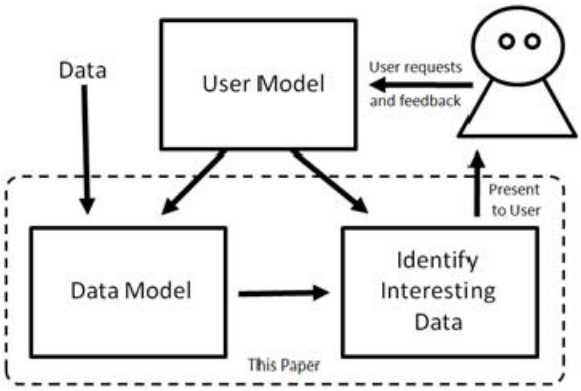

Figure 1. Components of an interactive machine learning solution

Most other papers on rare category detection also include mechanisms to include user feedback, and are therefore related to active learning methods. One of the first papers to use the term rare category detection [1] builds a mixture of Gaussian model with the expectation maximization algorithm, and then investigates a number of different ways to identify interesting samples. As users are presented these samples, they are asked for labels that identify which component the sample comes from. The mixture is then re-optimized with these labels at each iteration. In [6] the authors suggest a measure for local-density-differential-sampling that has similar motivations to our approach. They use a nearest neighbor type approach to identify samples next to sharp fall-offs in the density function. In [7] the authors use a similar metric in conjunction with a more sophisticated semi-supervised modeling technique. A more applied paper [8] presents the ALADIN network intrusion system for cyber-security. It uses a combination of active learning and rare category detection ideas, and demonstrates their utility in a practical problem. Authors in [9] develop an active learning sampling scheme for anomaly detection.

3. RARE CATEGORY DETECTION

In many real world data applications the aim is to detect, or to discover, unknown or unexpected events in massive data-streams. The dominant statistical learning problem in this kind of exploratory data analysis is anomaly detection: the data is modeled as a parametric or non-parametric density, $p(x)$, and low probability samples are presented to the user.

Rare category detection suggests that in many practical applications, we can say more about the data distribution. Specifically, it suggests that categories of anomalies will tend to form compact clusters in the input space, and that this will lead to regions where the density is greater than the local background [6]. It also suggests that as a consequence that

categories of anomalous samples are self-similar and therefore learning algorithms only need to find one representative example from each cluster to satisfy a user. This hypothesis is implemented in the synthetic data sets of Figure 2.

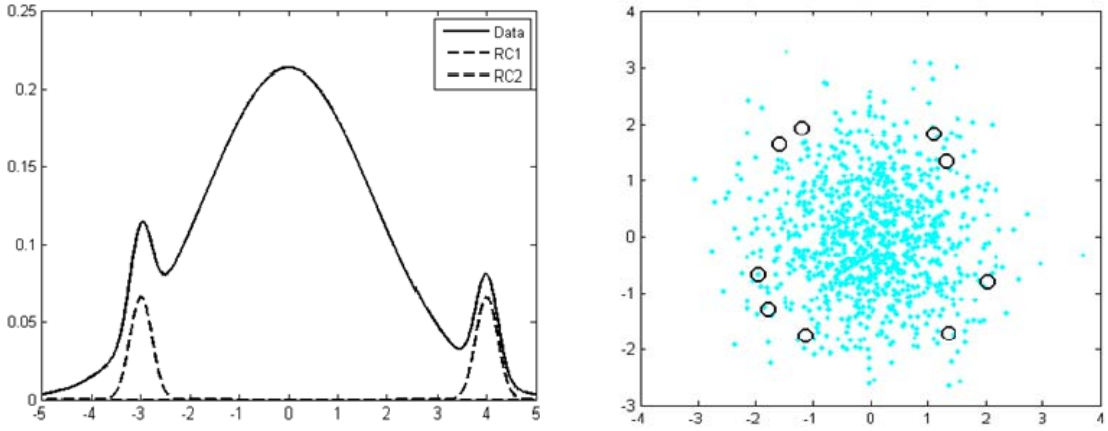

Figure 2. Left) A one dimensional mixture of Gaussians used to illustrate key ideas and Right) A two dimensional mixture used in synthetic experiments.

The one dimensional density on the left is used for illustrative purposes. The density is a mixture of Gaussians with three components. The largest background component contains 93% of the probability mass. Two other Gaussians are used to represent rare categories and each contributes 3.5% to the total mass. The two-dimensional density on the right is used in our synthetic experiments. In this case the background contributes 99% of the probability mass and the remaining 1% is distributed in 9 rare categories dispersed randomly with a bias towards lower probability regions of the background component. The objective of rare category detection is to choose a sample from each of the Gaussians. In the first case, this could theoretically be achieved in 3 samples, in the second case, 10 samples.

Given this objective it is informative to illustrate how different learning objectives prioritize (or weight) different data samples x. In figure 3 we show two different prioritizations of the data assuming the densities are known. On the left we show the weighting function for anomaly detection (scaled for display purposes):

$$w(x) = \frac{1}{p(x)} \tag{1}$$

This approach puts higher weights on low probability events, but it puts less weight on the rare category samples. On the right in Figure 3 we suggest a new objective for rare category detection:

$$w(x) = \frac{p(x)}{p_S(x)} \tag{2}$$

Where $p_S(x)$ is a smoothed version of the density:

$$p_S(x) \propto \int_{\acute{x} \in N(x)} p(\acute{x})\, d\acute{x} \tag{3}$$

That is, the smoothed density is proportional to the average probability in a local neighborhood $N(x)$. On the right in Figure 3 we see that this new ratio (which we call HighPass) gives higher weight to the rare categories, and that lower probability categories are given higher weight. A constant value of 1 has been subtracted from the HighPass curve for display purposes. Note the size of the neighborhood used in Equation 3 will affect what local peaks are given emphasis, and is a therefore an important free parameter with the proposed approach.

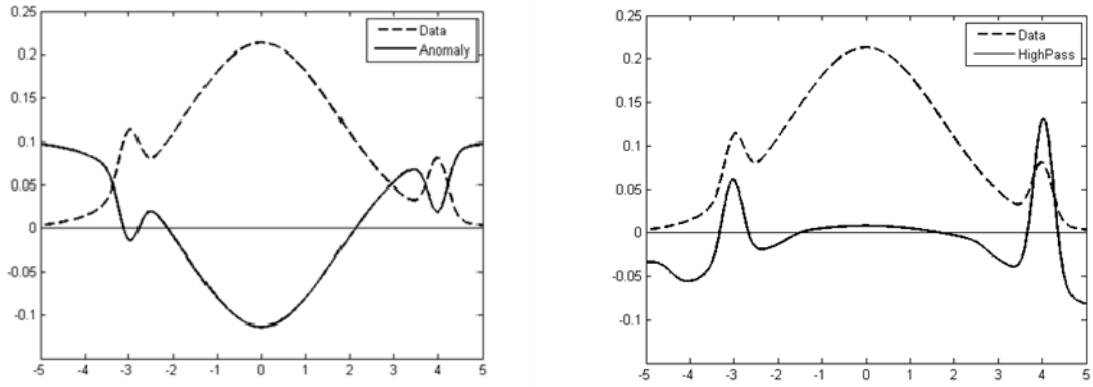

Figure 3. Left) Sample weighting function produced by Equation 1 (Anomaly) and Right) Sample weighting function produced by Equation 2 (HighPass).

3.1 Solution Methods

In practice the distributions are unknown and must be estimated from data. Two general approaches are Kernel Density Estimation (KDE) and Nearest Neighbor (NN) density estimation. For n data samples, Equation 2 can be estimated with:

$$KDE(x) = \frac{k/n}{l/n} = \frac{k}{l} \tag{4}$$

where k is the number of samples that fall within a distance d_1 of x and l is the number of samples that fall within a distance d_2 of x, where d_1 and d_2 are free parameters and we must choose the distance function (or kernel). In our experiments we use the Euclidean distance. For nearest neighbors we have:

$$NN(x) = \frac{d(x,k)}{d(x,l)} \tag{5}$$

where $d(x,k)$ is the distance from x to the k^{th}nearest neighbor, and where k and l are the free parameters. Figure 4 shows these estimates on the toy problem with 500 samples. We observed that while KDE typically produces smoother estimates, the nearest neighbor method appeared to have less variance at the tails of the distribution. Perhaps a more important difference is the parameterization. In the KDE approach we must provide two distance measures which correspond to expected distances between samples in the rare category and in the background density. In the NN approach we must provide two parameters related to the expected number of samples in the rare category and background density. Different applications may prefer one parameterization over another.

3.2 Synthetic Experiments

To illustrate the differences between anomaly detection and rare category detection we performed two sets of experiments. On the left in Figure 5 we evaluate algorithms against the more traditional criteria: the fraction of rare category samples that are selected compared to the number of samples drawn. On the right in figure 5 we evaluate algorithms against the rare-category detection criteria: the number of categories where at least one sample has been selected, compared to the number of samples drawn. We compare three algorithms: 1) Random sampling simply shuffles the data, 2) Anomaly detection fits a Gaussian to the data and sorts samples in ascending order of probability, and 3) HighPass sorts samples in descending order of weight, estimated with equation 4.

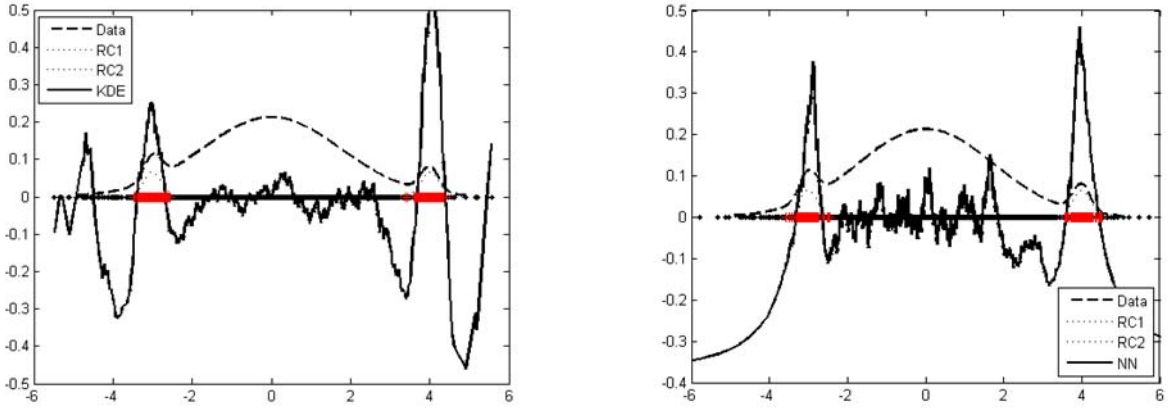

Figure 4. Left) A Kernel Density Estimate (Equation 4) with $d_1 = 0.1$ and $d_2 = 1$ and Right) Nearest Neighbor Estimate (Equation 5) with $k = 20$ and $l = 75$.

Evaluated against the traditional criteria, anomaly detection outperforms random sampling and HighPass by close to an order of magnitude. HighPass is able to quickly identify a fraction of the rare category samples, but has similar performance to random sampling if all samples are required. This result is consistent with the second experiment results, where only one sample is required from each category. In this case HighPass requires less than half as many samples as anomaly detection which in turn requires less than half as many samples as random sampling.

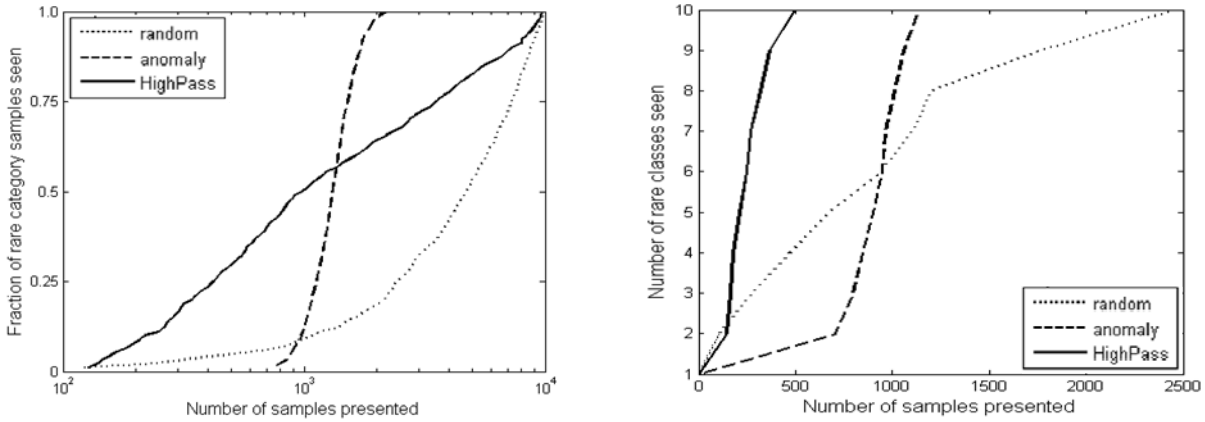

Figure 5. Left) Performance comparison when detecting all rare category samples and Right) comparison when detecting at least one sample from each category.

3.3 Solution via Classification

Density estimation is a difficult problem, particularly in the low density regions of the space. It is also possible that by dividing two density estimates, we are compounding the error [10]. In addition, the KDE and NN both have free parameters which must be somehow selected. An attractive solution to some of these problems is to cast the problem as a two class classification problem:

$$I(x) = \begin{cases} 1 & \frac{p(x)}{p_S(x)} > t \\ 0 & otherwise \end{cases} \tag{6}$$

Instead of producing a weighting function, classification produces an indicator function, $I(x)$, that divides the data into two sets: samples whose ratio is above a threshold are those that are not. Instead of producing a prioritized list, we

specify how many samples we want to show the user by choosing a threshold. In many applications this approach is a good match for users since they often have a fixed amount of effort available for the search task.

The solution method for Equation 6 has much in common with how machine learning casts anomaly detection as a classification problem [11]. In this case the indicator function looks like:

$$I(x) = \begin{cases} 1 & \frac{U(x)}{p_S(x)} > t \\ 0 & otherwise \end{cases} \quad (7)$$

where $U(X)$ is the uniform density. To cast this as a classification problem, a two-class training set is artificially generated. The first class (numerator) is a random set of points that have been sampled uniformly from the input domain, and given a class label +1, and the second class (denominator) is the data with class label -1. To produce the indicator function, this training data is simply provided to a standard classification engine, such as a Support Vector Machine (SVM). One of the advantages of this approach, is that we can use cross-validation and other model selection techniques to help choose the free parameters, such as the amount of regularization in the density estimates. On the left in Figure 6 we show a typical indicator function (scaled for display purposes) generated by this approach using a Gaussian kernel SVM.

We propose a similar approach can be used to estimate Equation 6. The first class in the artificial training set is the data. To generate the second class, we take each data sample and add uniform noise whose magnitude is proportional to the neighborhood size in Equation 3. On the right we show the typical output from this approach, again using a Gaussian kernel SVM. In the toy problem, the proposed classifier appears to being behaving as we would like. However in practical experiments we found using cross validation to select the SVM free parameters was non-trivial. This is partly due the highly skewed class distributions associated with rare category detection, which means the threshold in Equation 6 must be set (via class weights in SVM solutions) to produce very few false alarms. It is also related to the high variance that inevitably comes when estimating Equation 2 in low density regions (observe the left-most peak in Figure 6 Right). Addressing these issues is a topic of future research.

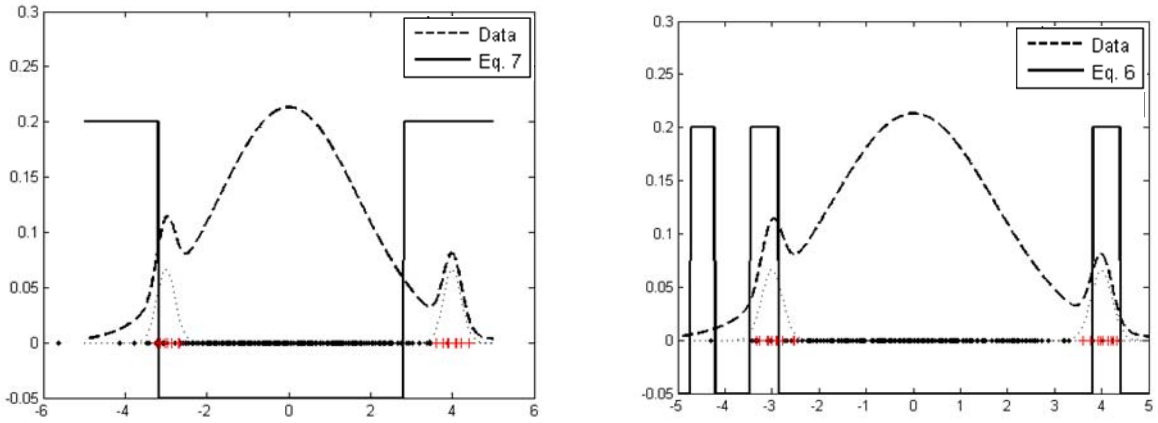

Figure 6. Left) Indicator function produced via anomaly detection using a SVM classifier and Right) Indicator function produced by Equation 6 using a SVM classifier.

3.4 Real-World Experiments

To evaluate the performance of the proposed approach with real data, we use the Shuttle and Abalone datasets from the UCI machine learning repository [12]. These datasets appear in many of the publications described in Section 2 since they are highly skewed mutli-class problems where a large fraction of the data belongs to a small number of *background* classes. In previous work these datasets are often sub-sampled (for computational reasons). In this paper we use the original datasets, and the distribution for the classes is illustrated in Figure 7.

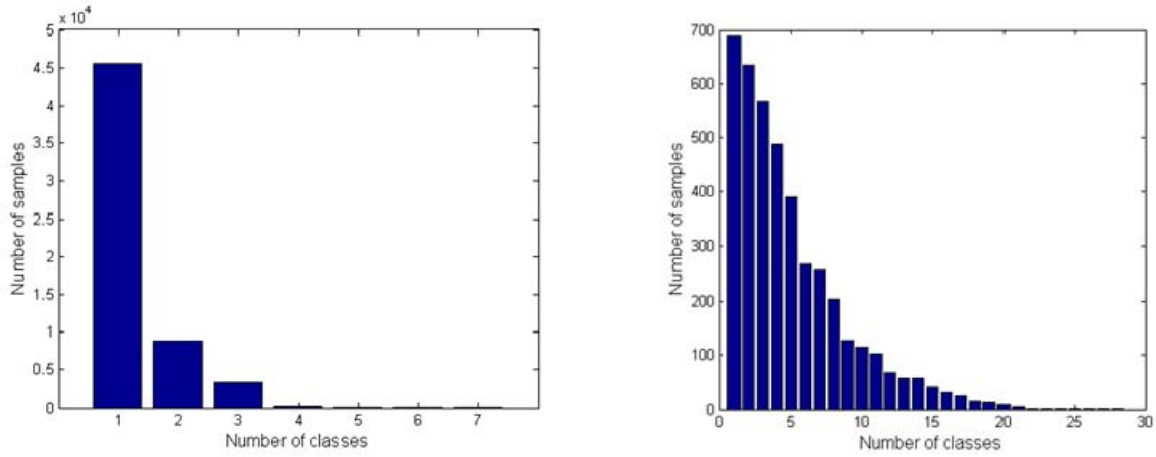

Figure 7. Left) Shuttle dataset class distribution and Right) Abalone dataset class distributions.

We compare the performance of three different methods: 1) Anomaly detection: as with the synthetic experiments, we fit a Gaussian to the data and then sort samples in ascending order of probability. 2) Interleave: an approximation of the technique presented in [1] and briefly described in Section 2. Our approximation (required due to the size of the Shuttle dataset) applied k-means clustering, instead of expectation maximization, and then *interleaved* the most distant samples within each cluster. 3) HighPass: estimated by Equation 4. Both Interleave and HighPass methods require the selection of free parameters. In this paper we perform an exhaustive search for the parameters that minimize the number of samples presented to observe one sample from all categories. The results are summarized in Figure 8.

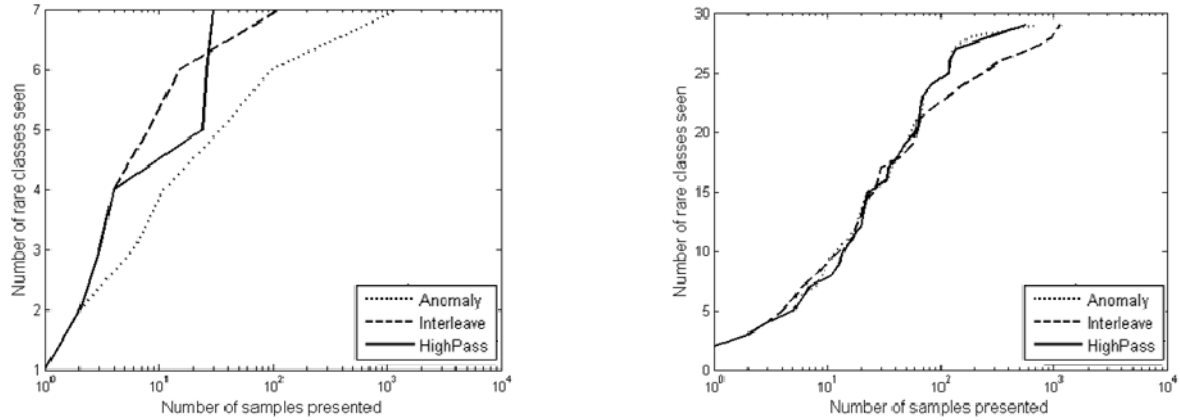

Figure 8. Performance comparison on Left) Shuttle dataset and Right) Abalone dataset.

On the left in Figure 8 we observe that both interleave and HighPass outperform anomaly detection by an order-of-magnitude. The optimal number of classes for the Interleave procedure was 200, far greater than the number of classes. On the right in Figure 8 we found the optimal parameter values for HighPass equated to anomaly detection. That is, the best performance was achieved by setting d_1 in Equation 4 very small so that the estimate only included 1 sample. This is consistent with the fact that a large number of classes in the Abalone dataset have very limited numbers of samples (e.g. 5 of the 28 classes only have 1 sample).

4. BROAD-AREA SEARCH IN REMOTE SENSING MAGERY

One application where interactive machine learning will be particularly useful is broad-area search of geo-spatial imagery. Large quantities of imagery are being collected from panchromatic wide-area motion imagery to multi- and hyper-spectral datasets. Content of interest is typically rare and ill-defined, but given enough time and resources, human operators "would know it if they saw it". Change detection provides an important framework for interactive search in this domain. Expert analysts can develop a very sophisticated understanding of geo-spatial environments, and change detection empowers them to keep up with the rapid rate of new data being accumulated over a particular geographic area.

Recently, a machine learning framework has been proposed which suggests a prioritization scheme for change detection [13]. Given two images, X and Y, this framework proposes the weighting function:

$$w(x,y) = \frac{p(x)p(y)}{p(x,y)} \tag{8}$$

We have developed a prototype interface for remote sensing imagery which is illustrated in Figure 9. It is based on Kitware Inc's Geospatial extensions to the Visualization Tool-Kit [14]. The key features of this interface include:

1. Three-dimensional navigation of geo-spatial data through pan/tilt/zoom controls similar to Google Earth and other geo-spatial data visualization environments.
2. Movie-like controls to control the refresh rate of images taken at different times.
3. A query mechanism that provides user cues based on Equation 8.
4. A simple interface that encourages users to annotate samples, providing feedback for updating prioritization schemes.

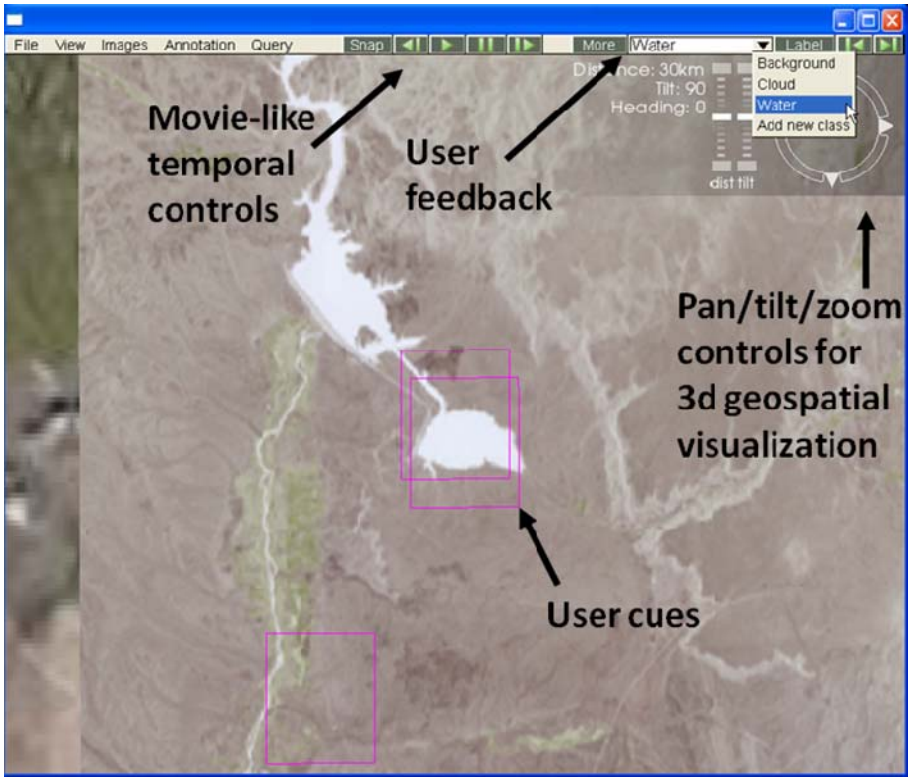

Figure 9. Prototype interface for identifying anomalous changes in remote sensing imagery.

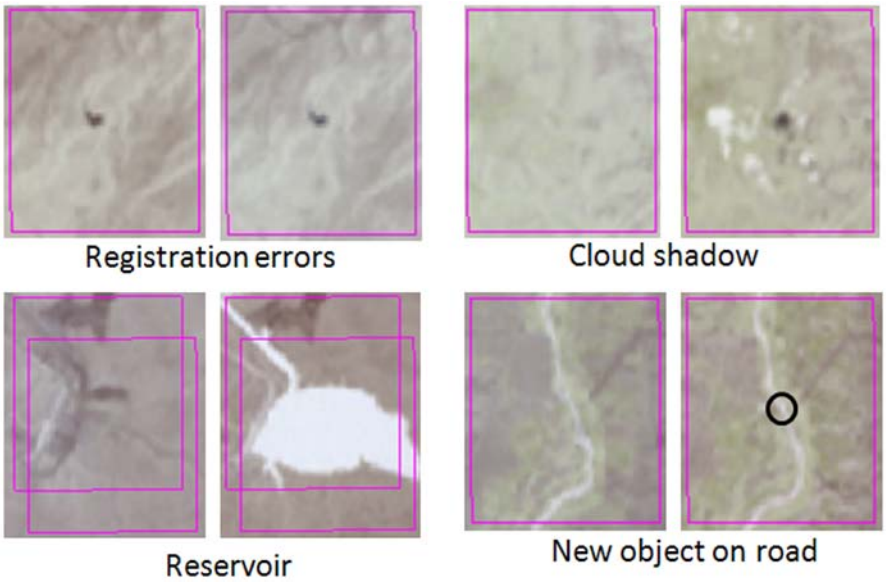

Figure 10. Examples of categories of change detected in 4-band imagery. For each example we show two images collected over a 10 year period.

In figure 10 we show some highly ranked locations identified by Equation 8. Many of these anomalous changes are uninteresting e.g. differences due to mis-registration and cloud shadow, but some may be very interesting to particular users, e.g. a small pixel-sized object appears near a road. When multiple instances of a category are identified in close proximity (e.g. the double hit on the Reservoir example) they can be easily discounted by a user. However in other cases (e.g. cloud shadows across an image) multiple instances can slow down and/or distract the user. In future work we will investigate if ideas presented in this paper can help address this problem.

5. SUMMARY

Rare category detection appears to be a good match for how people use machine learning algorithms because it assumes that a user will recognize a new category when they see it. This is difficult to guarantee in practice due to noisy or complicated data, and/or distractions in the exploitation environment. Our longer term research objective is to mitigate these problems and build rare category detection into a theoretical and practical machine learning framework for large-scale, user-in-the-loop data exploitation. While our approach is machine learning focused, it has much in common with the Visual Analytics research agenda. We hope to see increased interaction between these two fields in the near future to develop each of the following: new machine learning theory and algorithms that are robust to user decisions; new visualization environments that empower users to adapt interfaces as they explore and discover new information; and new mechanisms to capture user interactions to identify and reduce user uncertainty and predict user priorities.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the U.S. Department of Energy through the LANL/LDRD Program for this work. Thanks to NASA for allowing us to use the shuttle datasets.

REFERENCES

[1] D. Pelleg, and A. Moore, "Active Learning for Anomaly and Rare-Category Detection." Proc. 18th Annual Conference on Neural Information Processing Systems, (2004).

[2] C. R. Aragon, S. S. Poon, G. S. Aldering *et al.*, "Using visual analytics to maintain situation awareness in astrophysics." Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on, 27-34, (2008).

[3] W. A. Pike, J. Stasko, R. Chang *et al.*, "The science of interaction," Information Visualization, 8(4), 263-274 (2009).

[4] S. Fine, and Y. Mansour, "Active sampling for multiple output identification," Machine Learning, 69(2-3), 213-228 (2007).

[5] P. Vatturi, and W.-K. Wong, "Category detection using hierarchical mean shift." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France, 847-856, (2009).

[6] J. He, and J. Carbonell, "Nearest-Neighbor-Based Active Learning for Rare Category Detection." NIPS: Neural Information Processing Systems, Vancouver, B.C., Canada (2007).

[7] J. He, L. Yan, and R. Lawrence, "Graph-based rare category detection." Proceedings - IEEE International Conference on Data Mining, ICDM, 833-838, (2008).

[8] J. W. Stokes, J. C. Platt, J. Kravis *et al.*, [ALADIN: Active Learning of Anomalies to Detect Intrusions] Microsoft Research, MSR-TR-2008-24, (2008).

[9] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, 504-509, (2006).

[10] T. Suzuki, M. Sugiyama, J. Sese *et al.*, "Approximating Mutual Information by Maximum Likelihood Density Ratio Estimation." JMLR: Workshop and Conference Proceedings 4, 5-20, (2008).

[11] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," Journal of Machine Learning Research, 6, 211-232 (2005).

[12] A. Asuncion, and D. J. Newman, [UCI Machine Learning Repository], Irvine, CA(2007).

[13] J. Theiler, "Quantitative comparison of quadratic covariance-based anomalous change detectors," Applied Optics, 47, F12-F26 (2008).

[14] Kitware Inc., [Visualization ToolKit], (2010).