User-Driven Sampling Strategies in Image Exploitation

Neal Harvey, Reid Porter*
Intelligence and Space Research Division, Los Alamos National Laboratory.
Los Alamos, New Mexico, USA 87545

ABSTRACT

Visual analytics and interactive machine learning both try to leverage the complementary strengths of humans and machines to solve complex data exploitation tasks. These fields overlap most significantly when training is involved: the visualization or machine learning tool improves over time by exploiting observations of the human-computer interaction. This paper focuses on one aspect of the human-computer interaction that we call user-driven sampling strategies. Unlike relevance feedback and active learning sampling strategies, where the computer selects which data to label at each iteration, we investigate situations where the user selects which data is to be labeled at each iteration. User-driven sampling strategies can emerge in many visual analytics applications but they have not been fully developed in machine learning. User-driven sampling strategies suggest new theoretical and practical research questions for both visualization science and machine learning. In this paper we identify and quantify the potential benefits of these strategies in a practical image analysis application. We find user-driven sampling strategies can sometimes provide significant performance gains by steering tools towards local minima that have lower error than tools trained with all of the data. In preliminary experiments we find these performance gains are particularly pronounced when the user is experienced with the tool and application domain.

Keywords: visual analytics, interactive machine learning, active learning, relevance feedback

1. INTRODUCTION

Interactive machine learning is an emerging field of research that has similar aims to visual analytics: to leverage the complementary strengths of humans and machines to produce better solutions to data exploitation tasks. Perhaps the only real difference between interactive machine learning and visual analytics is historical: visual analytics has emerged from the visualization science community [1] and interactive machine learning has emerged from the machine learning community [2]. Visualization science has traditionally focused on the user, and has developed a number of tools and techniques that tailor user interfaces to the data exploitation problem, with the objective of maximizing user productivity. Machine learning has traditionally focused on the machine, and has developed a number of tools and techniques that tailor the data processing tools to the problem at hand, with the objective of maximizing prediction accuracy.

One of the main areas where interactive machine learning and visual analytics overlap is training: examples of tool inputs and outputs are used to tailor the tool to the application. In traditional machine learning, training examples are obtained in any number of ways, but in interactive machine learning, training examples are obtained from end-users in the deployed environment, as they interact with their data. This opens the door to a number of research questions for visualization science (e.g. how best to elicit training examples from end-users?) and for machine learning (e.g. how to best characterize user interactions in terms of training data?).

In Section 2 we describe recent machine learning advances that enable new forms of user interaction to be captured and incorporated into training processes. There are two main research thrusts to these training advances: 1) advances in the training vocabulary enable users to provide more information than standard labels and 2) advances in the training dialog enable users to interact in a more iterative, and intuitive way.

User-driven sampling strategies represent a new approach for the training dialog. These strategies emerge naturally in many interactive visual analytics applications, but they are yet to be formally developed in machine learning. In Section 4 we describe practical experiments that we use to quantify the potential benefits of user-driven sampling strategies. In Sections 5 we present our experimental results and discussion before concluding in Section 6.

* {harve, rporter}@lanl.gov

2. HUMAN-COMPUTER INTERACTION IN TRAINING

There are two main technical components that determine how human-computer interaction is translated into training data to build better machine learning tools. These are illustrated as axes of an interactive machine learning design space in Figure 1.

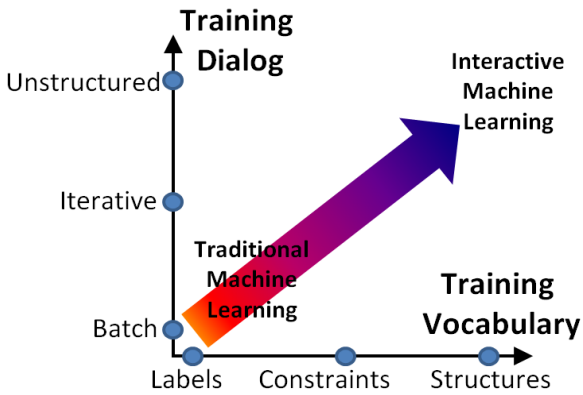

Figure 1.The design space for interactive machine learning in terms of training interactions.

We call the horizontal axis the Training Vocabulary. In traditional machine learning the vocabulary is based on simple labels. But over the last ten years, learning by example has advanced rapidly to include a much richer class of data-structures that can support a much richer set of user interactions.

A common application that exploits these more complex interactions is clustering. Typically, the interaction is formalized as equivalence constraints: pairs (or sets) of data that belong to the same cluster and/or pairs (or sets) that belong to different clusters [3]. These constraints can be obtained from the user through labeling interfaces, or, through *drag-and-drop* type interfaces where user's visualizing clusters are able to drag subsets of data closer to other subsets [4]. For example, in Bayesian Visual Analytics (BaVA) [5] the interaction is based on two- or three-dimensional point clouds, and users visually pick points and drag them closer to other points; this information is then used to refine the prior.

In the most general case, examples can be thought of as structures, or graphs [6]. These structures encode labels associated with subsets of data, but also relational (or semantic) relationships amongst different subsets of data. In general, structures are complex and collecting examples from users in an interactive setting is non-trivial. In addition, structures are typically fixed in advance and only approximate reality, which means generating training examples is often not intuitive. However, recent work in interactive machine learning has started to adapt these methods to interactive settings [7-9] and we suggest further work in this area can help exploit the spatial reasoning and semantic interactions that are inherent in many visual analytics systems [10].

2.1 The training dialog

The vertical axis is the Training Dialog, and it is the main focus of this paper. In traditional machine learning, training examples are collected up-front and provided to the training algorithm all at once (Batch learning). However, this is often not how users generate training data. Online learning methods relax the requirement that training examples be provided at the same time, but typically, online learning makes the same statistical assumptions as Batch learning: it assumes training samples are Independent and Identically Distributed (IID). This means (in principle) that the training data generated by users at time $t + 1$ should not be biased by the training samples provided previously, or the output from the machine learning system at time t. Relaxing this requirement has motivated a number of iterative learning techniques which are often a better match for how users want to interact with data. A very common dialog in iterative settings is relevance feedback [11] which is summarized in the following pseudo code:

```
1. Start with a small number of examples and build a content query.
2. Apply content query to unlabeled data and predict most relevant subsets.
3. User provides labels for predicted data indicating it is relevant (or not).
4. Update content query based on the new labels.
5. Goto 2
```

Active learning is very similar to relevance feedback, but it uses a different strategy for selecting examples (step 2). Active learning focuses on minimizing the number of labels required to obtain a given level of performance (the sample complexity). Note that with respect to the end-users application, these strategies may well select the most uninteresting samples in the data set. In some interactive applications this may not be a good match, however, in other applications, it can lead to better content queries with less work (labeling) for the user in the long term.

A long-standing challenge for relevance feedback and active learning is sampling bias. The samples that the computer selects in step 2 are not selected randomly, but the methods used in step 4 often assume they are. This means there are no guarantees that query performance will get better as more labels are obtained, and in fact, it may get worse. Mitigating sampling bias has been a key topic of research, and a number of methods have been developed that provide safety guarantees and batch learning performance in the worst case [12].

3. USER-DRIVEN SAMPLING STRATEGIES

So far we have discussed sampling strategies for step 2 where the computer determines which samples to label next based on the previous result. An alternative is to let the user choose which samples to label next. This approach is particularly relevant to visual analytics systems, since, in order for the user to choose which examples to label, they must be able to visualize (or browse) a larger subset of data. Empowering users to visualize and select the samples could have several potential advantages:

1. Users often know the most important aspects of the problem and can choose data appropriately.
2. By enabling users to interact with the sample selection and with predictions of the tool users can learn the strengths and weaknesses of the tool, and then choose examples that can guide the tool to better solutions.

For a concrete example of how user-driven sampling strategies emerge in interactive applications, we turn to image processing, and the task of labeling pixels. This application is the basis of the Crayons interactive machine learning system [13] as well as our own Genie image exploitation system [14]. This is illustrated in Figure 2. These tools obtain training examples from users through paintbrush-like tools. The insert in the lower left of Figure 2 shows a typical 'mark-up' where the user has selected examples of the feature of interest (vegetation) in green, and examples of the background in red. These labeled pixels are fed into a supervised learning method that produces a pixel-level classifier that can be applied to the entire image, as well as additional images. An example prediction from the pixel classifier is shown as a green / red overlay in Figure 2.

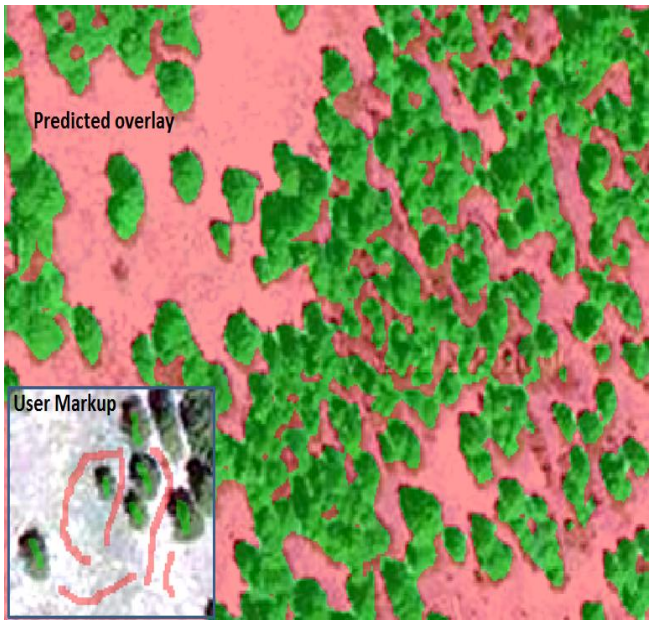

Figure 2. An interactive pixel classification tool that learns from user mark-up to predict features of interest.

This basic tool has a wide range of applications and our tool has been used in remote sensing, biomedical image analysis as well as material science. In many of these applications we have observed that users prefer to use the tool incrementally. Instead of providing a complete mark-up upfront, users provide a small amount of training data, train the classifier, then provide additional training data (typically where the classifier made a mistake) and so on. Informally, we have also observed that this iterative process often leads to better tools than those developed with batch training. In this paper we try to quantify and understand this potential performance improvement.

We note that the image classification problem is particularly well suited to user driven sampling strategies, because users can easily view a large amount of unlabeled data to choose samples. However, systems that support other applications and data types have also been developed [15], and we suggest understanding this interaction in more detail could potentially motivate application development in other visual analytics domains.

4. EXPERIMENTAL SETUP

To help quantify user-driven sampling strategies we performed a number of experiments with the image classification tool (Genie) described in Section 3. In each experiment the objective is to label each pixel in the image as feature (+1 / green) or background (-1 / red). Figure 3 provides an example. On the left is the input image (an 10 channel multi-spectral image) of a runway. The task is to separate the airplanes (feature) from the tarmac (background) and the task is

made concrete by the ground-truth mask shown in the middle of Figure 3. We show the ground-truth to the user at the start of the experiment to help the user understand the task objective, but it is not available to the user during the experiment. We also use the ground truth mask to measure performance at the end of each iteration.

4.1 IID Sampling

To provide a baseline for the experiment we train the classifier with training data selected randomly (IID) from the ground truth mask. In this paper we use Fisher's Linear Discriminant as the classifier in all experiments. On the right in Figure 3 is the performance of the classifier as the number of training examples is increased. We estimate the error by counting the number of mistakes made by the Fisher prediction compared to the ground truth mask (the Hamming distance between the predicted mask and the ground truth mask). We observe that the classifier has high bias on this problem: the classifier has non-zero error (the black line) even when the entire mark-up is provided to the classifier in training (the performance on the far right of the plot). We also observe the classifier has relatively low variance: the variation in error with different training sets is relatively small (roughly within $0.4x10^4$).

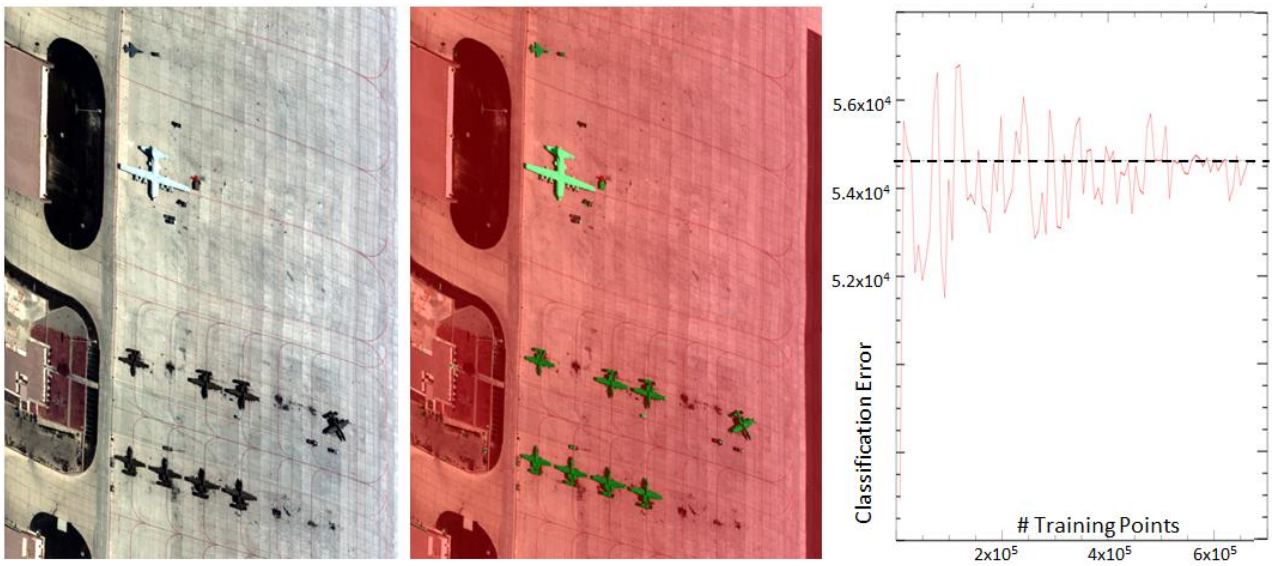

Figure 3. Left) RGB of a 10 channel multi-spectral image: Middle) Ground truth overlay showing desired features in green and background in red. Right) Performance of Fisher Linear Discriminant using IID sampling.

4.2 User-Driven Sampling

In the first iteration, the user inspects the raw image and selects a number of examples of feature pixels and background pixels. A typical selection for the aircraft problem is shown on the left in Figure 4. These examples are used to train a classifier which produces the prediction second from the left in Figure 4. We calculate the error of the prediction compared to the ground-truth. The second (and subsequent) iterations proceed much like the first, except now the user has the previous prediction to help them choose samples. We overlay the prediction with the image and the training data markup with different colors so that the user can simultaneously see the current training data and the current prediction. This typically biases the user's selection of samples towards misclassified pixels. This process continues until the user decides to stop (when they judge the prediction is no longer improving). In Figure 4, the user stopped after 9 iterations. The final training data used to build the classifier in iteration 9 is shown in black and white in Figure 4. The final prediction with this training data is shown on the right in Figure 4. Our user repeats this experiment several times. Each time, our user tried to start the initial markup on a different part of the image.

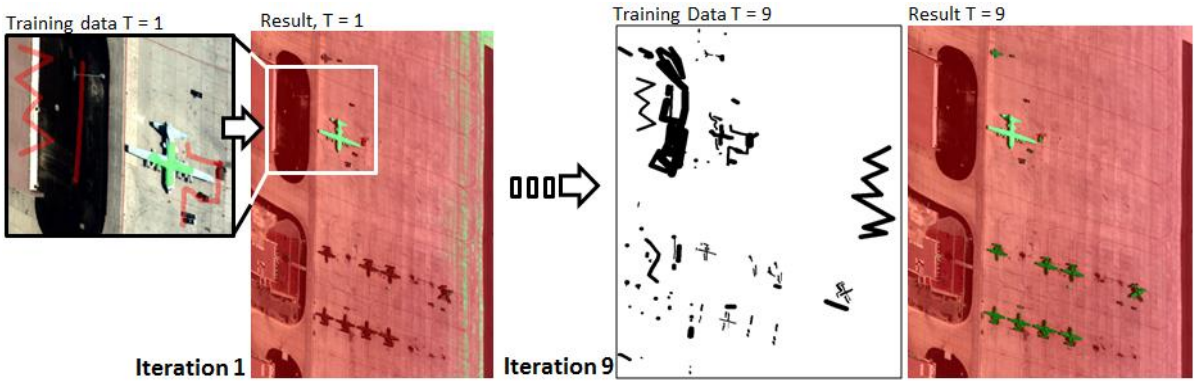

Figure 4. Far left) The user selects training samples in iteration 1: Second from left) the resulting classifier. Second from right) the labeled pixels after 9 iterations. Far right) the final classifier produced with these pixels.

5. EXPERIMENTAL RESULTS

5.1 Classifying Aircraft

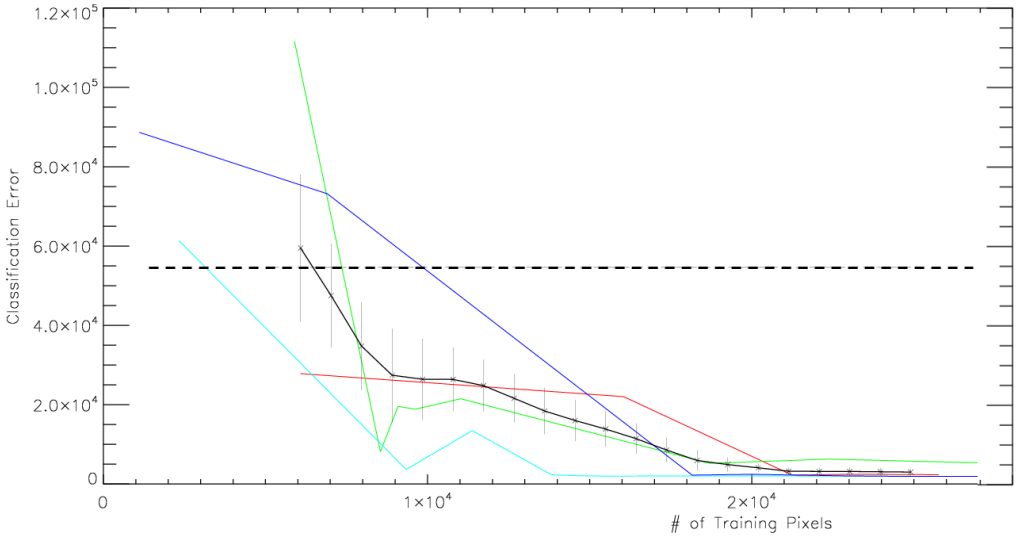

Figure 5. Performance as a function of training samples (iterations) for 4 different runs (colored) and the mean and standard deviation of errors (in black).

Figure 5 summarizes the experimental results for the aircraft problem described in Section 4. In the first iterations, the user's selection of samples appears to produce higher error than IID sampling. This implies the user's selection is biased compared to the final problem, as we might expect. However in subsequent iterations the user's sample selection leads to consistently better performance than IID sampling. This error is significantly lower than the variability in performance observed from IID sampling. The lowest number of mistakes observed in IID sampling was approximately 4.7×10^4, but user sample selection consistently resulted in predictions with less than 1×10^4 errors. The users sample selection is still biased, but it appears to be biased in a good way.

This result is somewhat different to relevance feedback and active learning results. In these results, the aim is to quantify how many samples are required for iterative learning algorithms to converge to the asymptotic IID error (the dashed line in Figure 5). The user-driven sampling results in Figure 5 tell a different story: user-driven sample selection appears to be driving classifier error below the asymptotic IID error. Note, if the user continues to iterate, they will eventually label the entire image, and the error curves in Figure 5 would return to the dashed line. This implies that users are stopping when the classifier is in a local minimum due to the biased training set.

5.2 Experiments using different features and image types

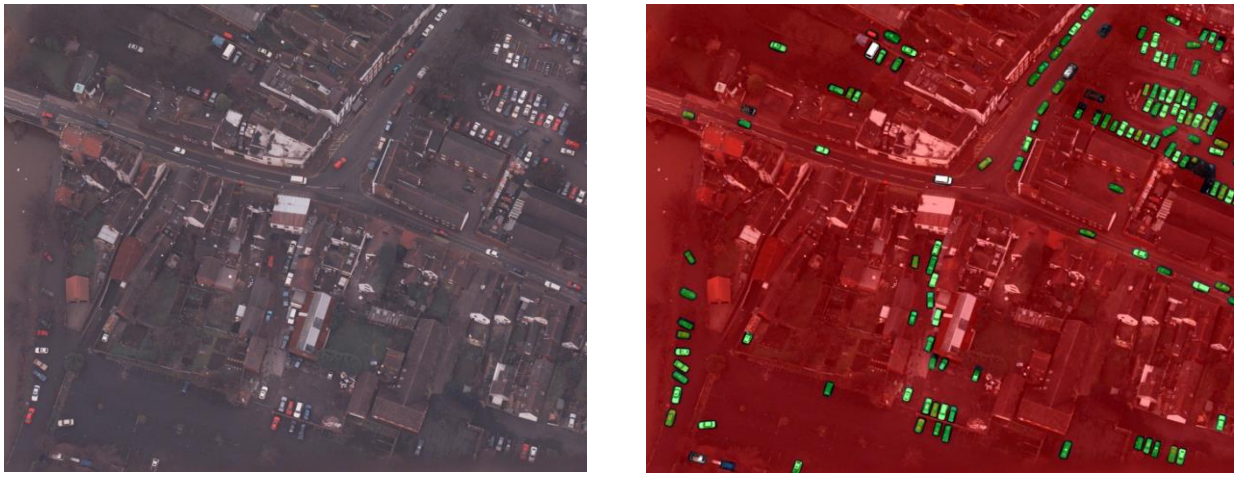

Figure 6. In this problem the image has 3 channels (Left), and the task is to delineate vehicles in an urban environment (Right).

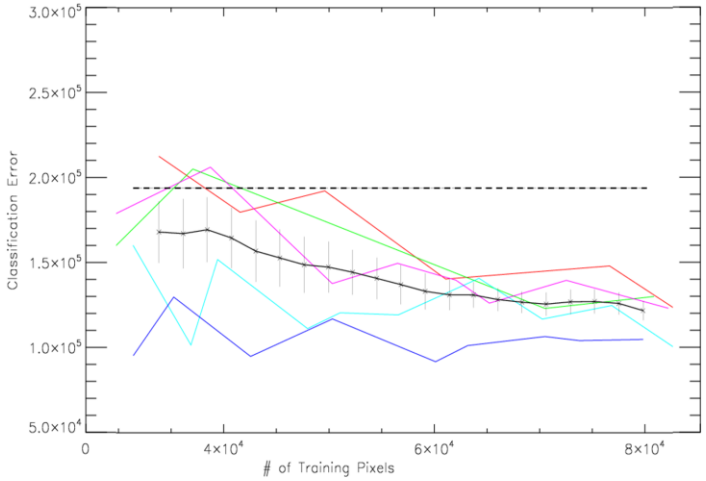

Figure 7. Performance as a function of iteration (training samples) for multiple trials on Figure 6.

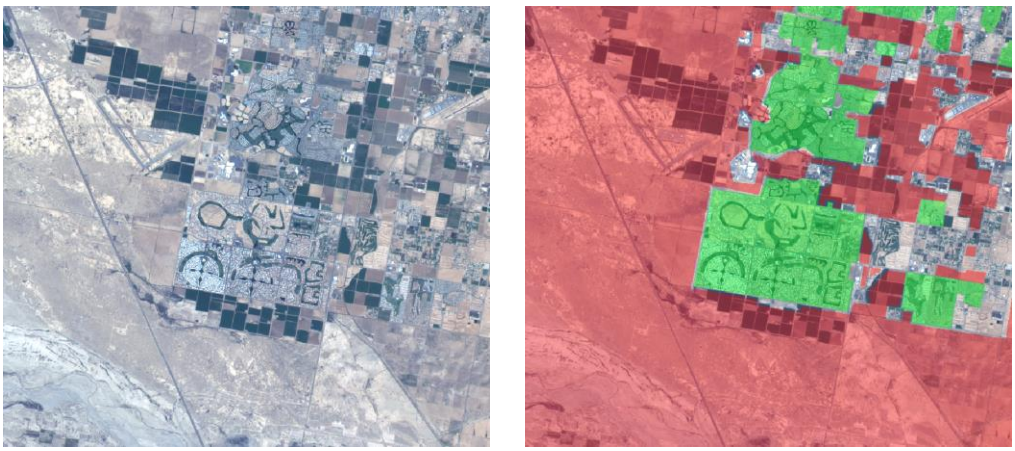

Figure 8. In this problem the image has 16 channels (Left), and the task is to delineate land use corresponding to high density residential neighborhoods.

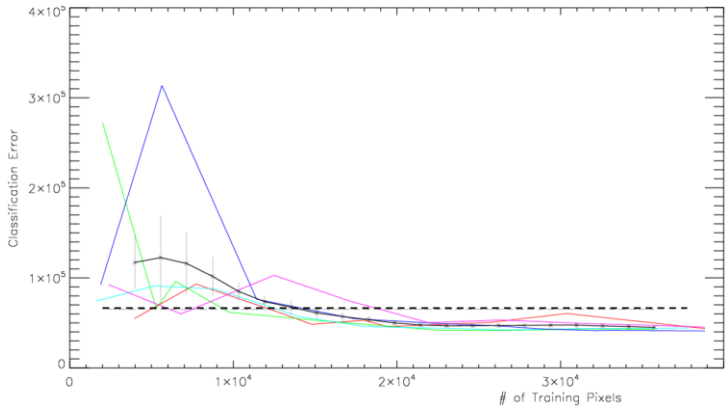

Figure 9. Performance as a function of iteration (training samples) for multiple trials on Figure 8.

We repeated the experiment on two other applications illustrated in Figures 6 and 8. In Figures 7 and 9 we observed a similar pattern of performance to the user sample selection experiment discussed in Section 5.1.

5.3 Experiment with an inexperienced user

In the experiments described so far the user involved was very familiar with the tool and the image analysis applications. In fact, through many years of hands-on experience, this user had already reached the conclusion that incremental mark-up could obtain better results than batch mark-up for these applications. To better understand the role of expertise in this situation, we solicited a second user, with no prior experience with the tool or image analysis. The results for the aircraft experiment (Figure 3) are summarized in Figure 10.

We observed that the inexperienced user was not able to reproduce the results of the experienced user in Figure 5, but they were able to consistently outperform the asymptotic error. Quantifying the performance as a function of experience (we have found that inexperienced users can quickly become familiar with our tool and applications) would be an interesting direction for future work.

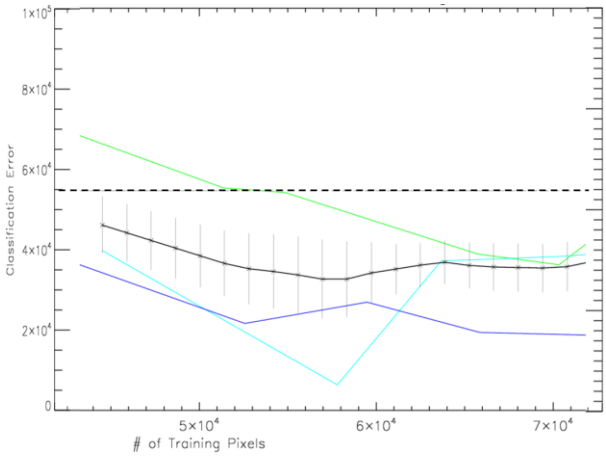

Figure 10. Performance of an inexperienced user as a function of iteration (training samples) for multiple trials on the aircraft delineation task in Figure 3.

6. SUMMARY

In this paper we identified a user-driven sampling strategy for machine learning that is often implicit in visual analytics systems. In preliminary experiments we found that user-driven sampling strategies led to biased training sets, but that this bias was often a good thing. One possible explanation is that as users iterate with the tool, they are learning the strengths and weaknesses of the classifier, and then choosing training samples that play to the classifiers strengths and mitigate its weaknesses. This of course is only possible due to the limitations of the classifier, and the classifiers used in

this paper all had high error (even when provided with all the training data). We hypothesize that if the classifier were better matched to the problem, then it would be harder for user-driven sample selection to improve upon the asymptotic error.

One of the ways user-driven sample selection impacts classifier design is through the class probabilities. We found that the number of samples selected by the user for each class in training was often very different to the class probabilities in the final prediction. Specifically, our applications typically had a large majority class (background), but in user-driven sample selection we observed class probabilities that were much more equal. This suggests user sample selection may be related to other methods in interactive machine learning such as ManiMatrix which enables users to interact with classifier design by adjusting weights on the classifier error matrix [16]. In the ManiMatrix system users interact directly with classifier design parameters, as they iterate with classifier predictions. In our approach, the interaction is indirect, but in some applications this could be more intuitive.

Our experiments also suggest a number of open questions for visualization science. The fact that our experienced user significantly outperformed our inexperienced user suggests a large number of visualization, training, and human factors could be involved in producing good user-driven sampling strategies.

REFERENCES

[1] J. J. Thomas, and K. A. Cook, [Illuminating the Path: The Research and Development Agenda for Visual Analytics] IEEE, (2005).

[2] R. Porter, J. Theiler, and D. Hush, "Interactive Machine Learning in Data Exploitation," IEEE Computing in Science and Engineering, 15(5), 12-20 (2013).

[3] S. Basu, I. Davidson, and K. L. Wagstaff, [Constrained Clustering: Advances in Algrotihms, Theory and Applications] Chapman & Hall / CRC, (2009).

[4] M. desJardins, J. MacGlashan, and J. Ferraioli, [Interactive Visual Clustering for Relational Data] Chapman & Hall / CRC, (2009).

[5] L. House, S. Leman, and C. Han, [Bayesian Visual Analytics: BaVA], (2010).

[6] L. Getoor, and B. Taskar, [Introduction to Statistical Relational Learning] The MIT Press, (2007).

[7] G. Zankl, Y. Haxhimusa, and A. Ion, "Interactive Labeling of Image Segmentation Hierarchies," Pattern Recognition, 7476, 11-20 (2012).

[8] L. Getoor, and C. P. Diehl, "Link mining: a survey," SIGKDD Explor. Newsl., 7(2), 3-12 (2005).

[9] R. B. Porter, S. Lundquist, and C. Ruggiero, "Learning to merge: a new tool for interactive mapping," Proc. SPIE, 87431F-87431F (2013).

[10] A. Endert, P. Fiaux, and C. North, "Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering," IEEE Transactions on Visualization and Computer Graphics, 18(12), (2012).

[11] Y. Rui, T. S. Huang, M. Ortega *et al.*, "Relevance feedback: a power tool for interactive content-based image retrieval," Circuits and Systems for Video Technology, IEEE Transactions on, 8(5), 644-655 (1998).

[12] S. Dasgupta, and D. J. Hsu, "Hierarchical Sampling for Active Learning," Twenty-Fifth International Conference on Machine Learning (ICML), (2008).

[13] J. A. Fails, and J. Dan R. Olsen, "Interactive Machine Learning," Intelligent User Interfaces, IUI '03, (2003).

[14] S. Perkins, J. Theiler, S. P. Brumby *et al.*, "GENIE - A Hybrid Genetic Algorithm for Feature Classification in Multi-Spectral Images," Proc. SPIE 4120 52-62 (2000).

[15] E. Zavesky, and S.-F. Chang, "CuZero: embracing the frontier of interactive visual search for informed users," Proc. 1st ACM International Conference on Multimedia Information Retrieval, 237-244 (2008).

[16] A. Kapoor, B. Lee, D. Tan *et al.*, "Interactive optimization for steering machine classification," Proc. 28th International Conference on Human Factors in Computing Systems, 1343-1352 (2010).