Examining Multiple Potential Models in End-User Interactive Concept Learning

Saleema Amershi[†], James Fogarty[†], Ashish Kapoor[‡], Desney Tan[‡]

[†]Computer Science & Engineering
DUB Group, University of Washington
Seattle, WA 98195
{ samershi, jfogarty }@cs.washington.edu

[‡]Microsoft Research
One Microsoft Way
Redmond, WA 98052
{ akapoor, desney }@microsoft.com

ABSTRACT

End-user interactive concept learning is a technique for interacting with large unstructured datasets, requiring insights from both human-computer interaction and machine learning. This note re-examines an assumption implicit in prior interactive machine learning research, that interaction should focus on the question "*what class is this object?*". We broaden interaction to include examination of multiple potential models while training a machine learning system. We evaluate this approach and find that people naturally adopt revision in the interactive machine learning process and that this improves the quality of their resulting models for difficult concepts.

Author Keywords

End-user interactive concept learning.

ACM Classification Keywords

H5.2 Information Interfaces and Presentation: User Interfaces. H1.2 Models and Principles: User/Machine Systems.

General Terms

Design, Human Factors, Performance.

INTRODUCTION AND MOTIVATION

Machine learning is a promising tool for enhancing human productivity and capabilities with large unstructured data sets. For example, consider a scientist trying to annotate segments of X-rays containing a specific "abnormality" in a medical imaging dataset or an office worker who wants a smart environment to automatically screen "unimportant" phone calls whenever it senses they are "busy". Interacting with individual objects to achieve these goals becomes difficult because of the vast amount of data (e.g., medical imaging archives or logs collected from sensing-equipped smart environments). With *end-user interactive concept learning*, people provide *examples* to interactively train a system to recognize concepts, such as "abnormality", "unimportant", or "busy". Automated processing is then based on those concepts. Because it can be challenging for an end-user and machine to reach a shared understanding of a concept, designing effective strategies for end-user interaction with machine learning is an important open challenge for human-computer interaction.

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.

This note re-examines an implicit assumption about how people should interact with machines that is common in prior interactive machine learning research (and this research thus complements research with an algorithmic focus, e.g., [2, 7, 10]). Machine learning systems learn by generalizing from examples of classes of objects. Prior work has thus focused interaction on prompting a person to answer "*what class is this object?*" [10, 7]. We propose that a better approach may be for a person to consider "*how will different labels for these objects affect the system in relation to my goals?*" Based on this approach, we examine end-user comparison of multiple potential models.

We situate this research in the context of CueFlik, a system that allows end-users to train visual concepts for re-ranking web image search results [5]. Consider a person attempting to train a "portrait" concept, as in Figure 1. Given a large and diverse set of images from a web query "Bill", the person may label Bill Gates and Bill Clinton images positive and dollar bill and Bill of Rights images negative.

Figure 1: This note considers the design of end-user interactive concept learning based on the insight that whether or not this magazine cover is a "portrait" may be less important than which resulting model is preferable.

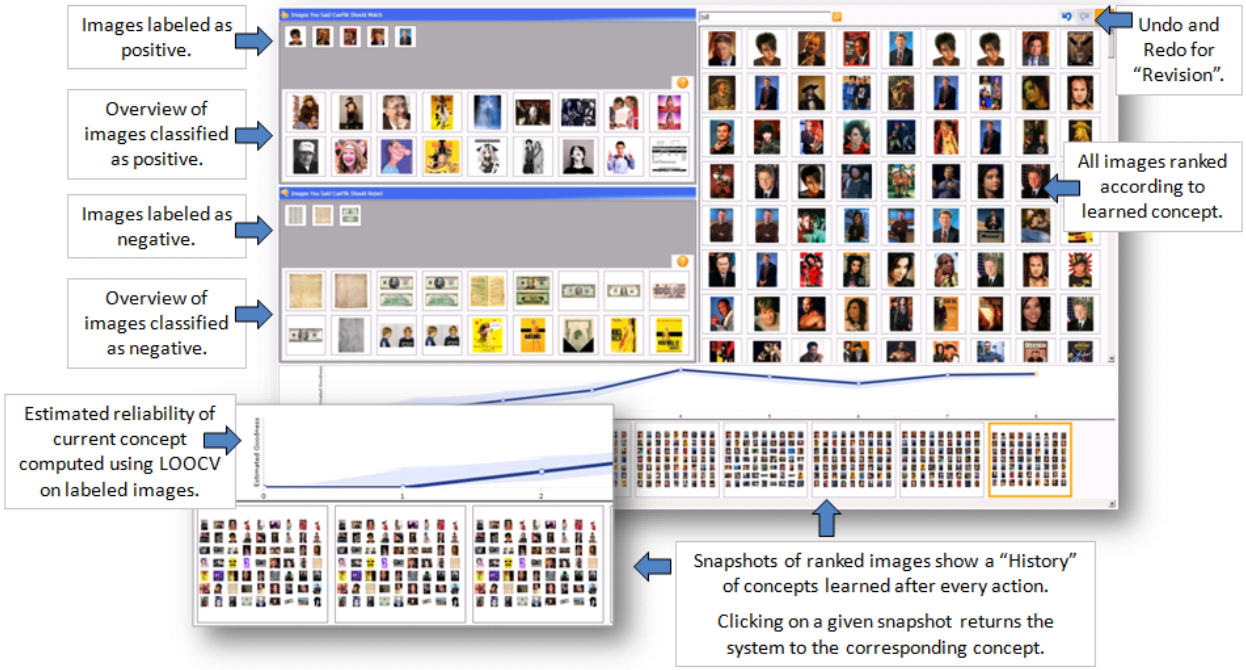

Figure 2: We investigate strategies examining multiple potential models while training CueFlik, a system for end-user interactive concept learning of visual concepts. Support for examining multiple potential models includes revision capabilities such as label removal, undo/redo, and a history visualization of recently explored potential models.

These are unambiguous given the person's goals. However, they may also encounter images they are less certain about, such as Bill Cosby on a magazine cover. The image features a prominent face, but also a variety of magazine graphics the person does not intend as part of the "portrait" concept. We propose a person should be able to experiment with either potential labeling, compare the resulting concepts, and then decide upon a label that guides the system to learn the desired concept. We examine this in CueFlik with (1) support for removing labeled examples directly and via undo/redo, and (2) a history visualization showing recently explored models, estimates of their reliability, and support for rolling back to previous models.

The specific contributions of this note are:

- A discussion of human-computer interaction and machine learning perspectives on the notion of examining multiple potential models. Bridging these perspectives suggests new strategies for end-user interactive concept learning.
- Discussion of our approach to implementing these strategies in CueFlik, as illustrated in Figure 2.
- An experiment evaluating our approach. We find that end-users naturally adopt revision as part of an interactive machine learning process and that undo/redo improves the quality of their resulting models for difficult concepts.

EXAMINING MULTIPLE POTENTIAL MODELS

Examining multiple alternatives is a proven technique in human-computer interaction research and practice. Supporting lightweight exploration of multiple alternatives has been shown to be effective in many tasks, such as complex multi-parameter problems in image editing [9]. Furthermore, supporting undo/redo is explicitly called out in Nielsen's ten usability heuristics [8].

In contrast, traditional interactive machine learning emphasizes collecting data to maximize information gain (e.g., [7, 10]). Interaction with a person has therefore generally been limited to asking "*what class is this object?*". Such an approach permits simulated experiments with fully-labeled datasets. However, treating a person as an oracle neglects human ability to revise and experiment. Deleting training examples, whether explicitly, by undo, or by rolling back to a previous model, is a poor fit from this perspective because it deletes information that has already been made available to the machine learning system.

Prior research at the intersection of human-computer interaction and machine learning has implicitly assumed the machine learning perspective. Fails and Olsen discuss the interactive machine learning process as providing additional training examples to improve the current classifier [4], but never raise the possibility of providing *different* training examples to improve a model. Dey *et al.* discuss iteratively providing demonstrations and annotations [3], but do not explore the possibility that performance might be improved by *different* examples. This note considers end-user interactive machine learning from a revision approach, wherein a person assigns different combinations of labels over time as they examine and choose from multiple potential models. General machine learning tools support comparison of different model parameterizations for a given set of data (e.g., Weka [11]), but do not support

comparing models as they evolve during interactive training. Furthermore, such tools target people skilled in statistical machine learning rather than the general end-user.

EXAMINING MULTIPLE POTENTIAL MODELS IN CUEFLIK

CueFlik is a system that allows end-users to train visual concepts for re-ranking web image search results [1, 5]. End-users train CueFlik by providing examples it should match (five images in Figure 2 upper-left) and examples it should reject (three images in Figure 2 middle-left). CueFlik uses these examples to learn a distance metric based on a set of visual features, which it then applies with a nearest-neighbor classifier to re-rank images. CueFlik guides end-users to provide informative examples using overviews of the current positive and negative regions of a learned concept (eighteen positive overview images and eighteen negative overview images in Figure 2 left) [1]. It also combines these overviews with a presentation of the entire set of images (labeled and unlabeled) ranked by their likelihood of membership in the positive class (seventy-two images visible on Figure 2 right) [5]. We made two sets of enhancements to CueFlik in this research.

History of Potential Models

We augmented CueFlik to include a history visualization of recently examined potential models (Figure 2 bottom). The history contains a plot of each model's estimated reliability, updated after each addition or removal of examples. Model reliability is measured using leave-one-out-cross validation on the current set of training examples, with confidence intervals computed according to a binomial distribution [6]. Estimating reliability via training data is necessary because standard measures (e.g., precision or recall) require labeled data (which does not exist in our scenario). The history also contains snapshots of the top ranked images for each model. The history is intended to help people visually compare and assess the relative quality of the models they have trained.

Model Revision

CueFlik includes several mechanisms for people to revise their currently trained model. First, a person can undo or redo actions. Second, they can remove labels of individual examples if they feel the example could be hurting the model's performance. Finally, a person can click directly on a data point or a snapshot in the plot to revert back to that model. The person can then continue providing examples from that stage, effectively enabling a simple branching mechanism for exploring multiple models.

EXPERIMENT

We conducted an experiment to understand how people would use our history and revision enhancements to examine multiple potential models in CueFlik and to determine the effectiveness of a revision-based approach to interactively training a machine learning system. We used a 2 (*History* vs. *No History*) × 2 (*Revision* vs. *No Revision*) within-subjects design. Conditions were counterbalanced using a Latin Square. Nineteen participants (8 female, ages 18-45) volunteered for the study (a twentieth was unable to attend as scheduled).

Participants trained three models in each condition, corresponding to concepts such as "pictures with products on a white background" and "portraits of people". CueFlik automatically issued a query for each concept (e.g., "drink", "bill") to obtain a diverse set of 1000 images previously retrieved from the web. Participants were then given a sheet of paper with ten target images on it and were asked to train the system to re-rank the images such that those like the targets are ranked highly (the target images themselves were removed from the set). Based on prior experience with these target concepts, we categorized them as easy or difficult and pseudo-randomly selected queries such that the first task in each interface was easy (for the purpose of practicing with the new interface) and the next two were difficult. We fixed the order of the queries selected because we did not expect them to lead directly to carryover effects. Participants were asked to train the system as accurately and quickly as possible and we imposed a maximum time limit of four minutes for each task. All actions were automatically logged and timestamped.

After each condition, participants were given a short questionnaire about the interface they had just used and the models they trained with it. At the end of the study, a final questionnaire collected overall assessments of CueFlik. The experiment lasted approximately 90 minutes, and participants were given a software gratuity for their time.

RESULTS

Analysis of logged data from our experiment shows that participants made use of the revision mechanisms to explore multiple potential models in CueFlik. When the history visualization was available, participants made revisions in 68% of their tasks. To make revisions, participants used the undo/redo feature in 19% of tasks, the remove label feature in 5% of tasks, and the ability to rollback to previous models via the history in 42% of tasks. When the history visualization was not available, participants made revisions in 41% of tasks. Interestingly, their usage of the undo/redo and remove label features increased to 30% and 11%, respectively, likely because these were the only revision mechanisms that were available in this condition.

For the tasks in which people made revisions when the history was available, 3% of their total actions (labeling examples and revision of any kind) were undo/redo actions, 1% were removing labels, and 9% were rolling back via the history. Without history, they relied more on the undo/redo and remove label features, using undo/redo in 11% of actions and removing labels in 3% of actions. This suggests participants were able to make progress by providing CueFlik with examples, but in some cases felt it necessary to explore or revert back to previous models.

To analyze the impact of our revision and history enhancements on participant ability to effectively train a concept, we measured their *Time* to complete each task, the *NumImages* and *NumActions* taken to train their final

models, and their final model *Scores*. Score is defined as the mean ranking of the target images by the final learned concept, where a lower score indicates a higher-quality concept (i.e., targets are nearer the top of ranked examples). We perform these analyses using mixed-model analyses of variance. All of our models include *History* (*History* vs. *NoHistory*), *Revision* (*Revision* vs. *NoRevision*), and their interaction *History*×*Revision* as fixed effects. To account for any variation in individual performance, query difficulty, or other carryover effects, we include *Participant* and *Query* as random effects. We exclude easy concepts because these were intended for practice with each condition and because we expect our enhancements to be less relevant in situations where there is little ambiguity.

Analyses reveal that participants spent more *Time* creating models with *History* available (160 vs. 146 seconds, $F_{1,118}$=3.93, p = .049) and performed more *NumActions* with *History* than without (17.2 vs. 15.1, $F_{1,118}$ =4.55, p = .035). There was no effect of *History* on *Score*. Participants also performed more *NumActions* when *Revision* capabilities were enabled compared to *NoRevision* (17.2 vs. 15.2, $F_{1,118}$=4.08, p = .046). There were no effects of *Revision* on *Time* or *NumImages*. In terms of final *Scores*, we found that participants created better models with *Revision* (211 vs. 242, $F_{1,119}$ =3.57, p = .061). There were no interaction effects on any of our dependent measures.

DISCUSSION AND CONCLUSION

Our evaluation shows that participants made use of revision mechanisms while interactively training a machine learning system and that this led them to achieve better final models in the same amount of time (even while performing more actions). Furthermore, being able to examine and revise actions is consistent with how people typically expect to interact with applications. One participant commented that without revision "it felt a little like typing on a keyboard without a backspace key".

In contrast, our history visualization enhancement led participants to spend more time and perform more actions to train concepts without improving overall model quality. While some participants seemed to find the history helpful for examining different models (e.g., "[the visualization was] helpful to see if I was heading in the right direction"), observations during the study and other participant comments indicate that the plot was generally distracting (e.g., "I felt like I was concentrating too much on the line graph."). Although the plot used an accepted machine learning metric to estimate model reliability (leave-one out cross validation accuracy), end-users seemed to use it less like an approximation tool for helping them interpret model quality and more like a quantity to maximize (e.g., "I wanted the graph to go up instead of concentrating on [the ranked results]"). Participants did, however, use the history for reverting back to previous models, suggesting that the history may be beneficial as a facility for enabling revision (e.g., "[the history] helped because when I started to get off track I could always go back and try a different route").

This research re-examines a traditional interactive machine learning focus on the question "*what class is this object?*" and broadens the interaction to include examining multiple potential models. Without such support, our study participants found it difficult to recover when model quality appeared to drop (e.g., "when I felt I was creating a good [model], sometimes it would get worse, and I had a hard time re-teaching it"). Furthermore, prior research has found that, after a certain point in the interactive machine learning process, continuing to provide examples can become detrimental to model accuracy [1]. Our research shows that including revision mechanisms can improve end-user interactive training of machine learning systems.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under grant IIS-0812590.

REFERENCES

1. Amershi, S., Fogarty, J., Kapoor, A., and Tan, D. Overview-Based Example Selection in End-User Interactive Concept Learning. *Proceedings of the ACM Symposium on User Interface Software and Technology* (UIST 2009), 247-256.
2. Beygelzimer, A., Dasgupta, S., and Langford, J. Importance Weighted Active Learning. *Proceedings of the International Conference on Machine Learning* (ICML 2009), 49-56.
3. Dey, A.K., Hamid, R., Beckmann, C., Li, I. and Hsu, D. a CAPpella: Programming by Demonstrations of Context-Aware Applications. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI 2004), 33-40.
4. Fails, J.A. and Olsen Jr., D.R. Interactive Machine Learning. *Proceedings of the ACM Conference on Intelligent User Interfaces* (IUI 2003), 39-45.
5. Fogarty, J., Tan, D., Kapoor, A., and Winder, S. CueFlik: Interactive Concept Learning in Image Search. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI 2008), 29-38.
6. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the International Joint Conference on Artificial Intelligence* (IJCAI 1995), 1137-1143.
7. MacKay, D.J.C. Information-Based Objective Functions for Active Data Selection. *Neural Computation 4*, 4 (1992), 590-604.
8. Nielsen, J. *Heuristic Evaluation*. In Nielsen, J. and Mack, R. L. (eds.). Usability Inspection Methods, John Wiley & Sons, New York, NY, 1994.
9. Terry, M. and Mynatt, E.D. Side views: persistent, on-demand previews for open-ended tasks. *Proceedings of the ACM Symposium on User Interface Software and Technology* (UIST 2002), 71-80.
10. Tong, S. and Chang, E. Support Vector Machine Active Learning for Image Retrieval. *Proceedings of the ACM Conference on Multimedia,* (2001), 107-118.
11. Witten, I.H. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition)*, Morgan Kaufmann, San Francisco, CA, USA, 2005.