Beyond Feature Relevance: Incorporating Rich User Feedback Into Interactive Machine Learning Applications

Krzysztof Z. Gajos
Harvard School of Engineering and Applied Sciences
kgajos@eecs.harvard.edu
http://www.eecs.harvard.edu/∼kgajos/

The success of interactive machine learning systems depends both on the machine and on the human performance. An understanding of machine capabilities and limitations should inform interaction design, while the abilities, preferences, and limitations of human operators should inform the choice of inputs, outputs, and performance requirements of machine learning algorithms. A relevant example from our past work is our Arnauld system [3] for active preference elicitation. A lot of previous work in that area solicited user feedback in the form numerical ratings over possible outcomes. However, unless the rating scale is well grounded, people tend to be inconsistent and unreliable providing this type of feedback. What works much more robustly is pairwise comparison queries, where the person only has to state which of two possible outcomes he or she prefers [1]. Adopting this input interaction, however, required us to develop a new learning algorithm. In turn, to account for the limitations of the algorithm, we implemented the example critiquing interaction [4] to allow people to manually direct the learning once the active learning process no longer resulted in rapid improvements in the model quality.

Currently we are working on incorporating richer user feedback into interactive machine learning systems. Typically, machine learning algorithms only solicit labels from the users but several projects (e.g., [2, 5, 6]) have shown that incorporating richer feedback—that captures the user's rationale—leads to faster and more generalizable learning. So far, this feedback has been limited to feature relevance. Is this the best or the only type of rich feedback we can elicit from users?

We have conducted a preliminary study in the context of preference elicitation for an e-commerce application to understand what types of feedback people naturally provide, and what the value of these different types of feedback might have for the speed and quality of learning. Specifically, we asked users to answer a set of pairwise comparison questions regarding digital cameras and we recorded their choices as well as free form explanations of their choices. We then manually analyzed those responses and we identified the following types of explanations:

- User indicated a relevant feature (”I hate AA batteries”)
- User indicated a relevant conjunction of features (”I love Nikon SLRs” — brand and camera type were encoded as separate features)
- User indicated a new feature that was only available in the textual description (”[I like the] touch screen”)

- User indicated a threshold for a numerical feature (”one is too cheap, the other too expensive”, ”this one has lower resolution than what I want”)
- User specified a general *ceteris paribus* preference (”I like Nikon over Canon”, ”I prefer an SLR over a compact camera”)
- User considered a hypothetical outcome (”I would choose Pentax for large touch screen if it cost the same as the other camera”)
- User indicated an explicit trade-off (”Resolution and size matter but the price is too steep”, ”I’d be willing to sacrifice AA batteries for the $250 price difference”)

We next conducted an experiment where we incorporated all these different types of feedback into the training data using the pseudo-document approach [5]. Conjunctions, thresholds, preferences and hypotheticals each modestly but consistently resulted in improved predictive accuracy (more so than identifying relevant features). Pseudo documents are an intuitive but relatively ineffective way to incorporate feedback on features [5]. We expect to see stronger results when we rerun the experiment using other approaches.

Given these pilot results, we plan to conduct a more formal study that will include several different types of learning problems in a few different application scenarios to give us a better sense of the generalizability of these findings. Our results so far already suggest that there may be even more useful types of rich feedback than just feature relevance. These results will impact both the algorithm and the interaction design for interactive machine learning systems.

References

[1] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there: Preference judgments for relevance. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 1532–1536, 2008. To appear.

[2] Gregory Druck, Gideon S. Mann, and Andrew Mccallum. Learning from labeled features using generalized expectation criteria. In Sung H. Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat S. Chua, Mun K. Leong, Sung H. Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat S. Chua, and Mun K. Leong, editors, *SIGIR*, pages 595–602. ACM, 2008.

[3] Krzysztof Gajos and Daniel S. Weld. Preference elicitation for interface optimization. In *Proceedings of UIST 2005*, Seattle, WA, USA, 2005.

[4] Pearl Pu and Li Chen. User-involved preference elicitation for product search and recommender systems. *AI Magazine*, 29(4), 2009.

[5] Hema Raghavan and James Allan. An interactive algorithm for asking and incorporating feature feedback into support vector machines. In *SIGIR ’07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Jul 2007.

[6] O Zaidan, J Eisner, and C Piatko. Using “annotator rationales” to improve machine learning for text categorization. In *Proceedings of NAACL-HLT*, pages 260–267, 2007.