Are Good Explainers Secretly Human-in-the-Loop Active Learners?

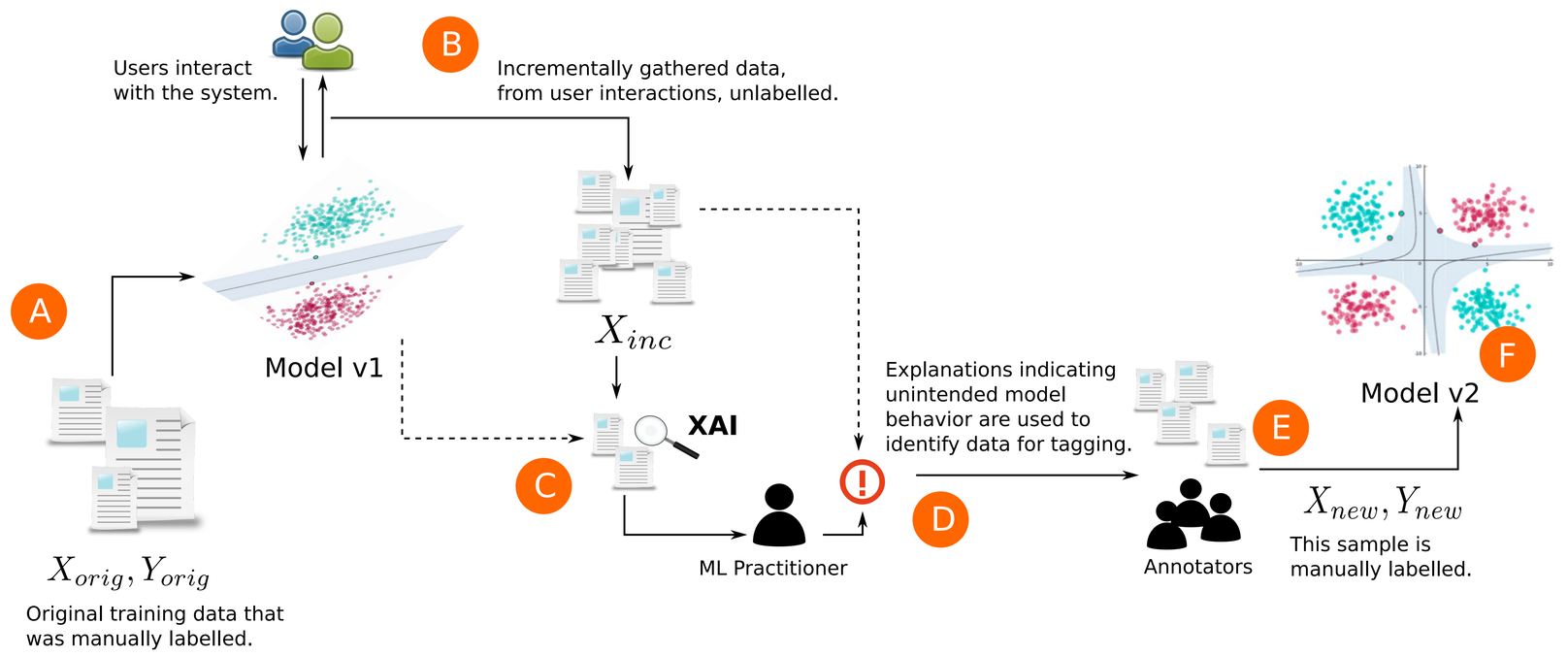
Emma Thuong Nguyen, Abhishek Ghose
[24]7.ai, California, USA
emma.nguyen@247.ai, abhishek.ghose@247.ai
Presented at **AI & HCI Workshop** at **ICML 2023**, Honolulu, Hawaii, USA

Key Takeaways

Study the use of Explainable AI (XAI) in selecting additional training data:

- Argue that this process is equivalent to Active Learning (AL) with a human-in-the-loop (HIL) query strategy.
- Mathematically approximate the role of the HIL and the end-to-end workflow for easy simulation.
- Present promising initial results.

Use Case



XAI-based Data Selection is Active Learning

Components effectively form a query strategy in AL:

- An explainer used to detect surprising patterns in model predictions.
- Explanations are then used to solicit further instances from an unlabeled pool of data.

Experiments

- Use *SHAP* Partition explainer and Bayesian Optimizer (*Optuna* within *Ray Tune*) to explore the search space of the explanation parameters. We employ *Support Vector Machines* with a linear kernel and *Universal Sentence Encoding* for classifying and representing text, respectively.
- Compare against some standard AL query strategies (we use the *modAL* library), such as sampling at random or based on entropy and margin.

Future Work

- Further comparisons with recent AL techniques, and additional explainers, classifiers and datasets.
- Use XAI to formulate other query strategies that capture some surprising model behavior.
- User studies to validate the approximation for human-in-the-loop process.
- *Learn* a query strategy exploiting the fact that our formulation of the workflow (Equation 3) is differentiable.

Mathematical Formulation

Step C

Explanations are sought for instances $x_i \in X_s$, where $X_s \subseteq X_{inc}$ is a set of N_s instances randomly selected from the unlabeled pool X_{inc}. For each instance $x_i \in \mathbb{R}^d$, an explanation weight vector $q_i \in \mathbb{R}^{d'}$, corresponding to $x'_i \in \mathbb{R}^{d'}$ in the explanation space, is produced.

Step D

Representing* unintended *model behavior: Different labels are predicted for a pair of instances that are either similar or produce similar explanations.

$$A_{ij} = (q_i \odot x'_i) \cdot (q_j \odot x'_j)^T \tag{1}$$

$$B_{ij} = \begin{cases} 1 & \text{different predictions for } x_i, x_j \\ 0 & \text{otherwise} \end{cases}$$

$$C = (A \odot B)\mathbb{1}_{N_s} \tag{2}$$

where $\mathbb{1}_{N_s} \in \mathbb{R}^{N_s \times 1}$. Intuitively, $C \in \mathbb{R}^{N_s}$ indicates preference weights for instances in X'_s.

***Retrieving* top *instances* X_{top} *from* X_{inc}** based on dot-product similarity to $x'_i \in X'_s$ augmented with preference weights:

$$W = X'_{inc}(C\mathbb{I}^T_{d'} \odot X'_s)^T \mathbb{1}_{N_s} \tag{3}$$

Steps E and F

Obtain labels Y_{top} corresponding to X_{top} via human annotation, and construct the following new dataset to retrain the model:

$$X_{new} = \begin{bmatrix} X_{orig} \\ X_{top} \end{bmatrix} \qquad Y_{new} = \begin{bmatrix} Y_{orig} \\ Y_{top} \end{bmatrix}$$

References & Paper