PURDUE UNIVERSITY
Department of Computer Science

Give Weight to Human Reactions: Optimizing Complementary AI in Practical Human-AI Teams

Hasan Amin, Zhuoran Lu, Ming Yin

Overview

- **Improve performance in human-AI (HAI) teams by designing AI-driven decision aids that *take humans' reaction when interacting with it into consideration*.**
- **Highlights:**
  - Focus on adjusting AI models based on humans' confidence in their own decisions.
  - Derive optimal training scheme under assumed, threshold-based team decision making model.
  - Validate efficacy in practice through systematic experimentation on synthetic and real-world data
- **Our (O) approach complementary to most existing (E) methods:**
  - E: adjust humans to better utilize given AI
    O: *adjust AI to better benefit human teammate*
  - E: design AI for maximum individual accuracy or simulate assumed behavior without 'actual AI'
    O: *train actual AI for complementarity and team gains*
  - E: rely on assumption that humans are rational
    O: *build on lack of rationality and human biases*

Problem Setup

- Human-AI (HAI) joint decision making setting, where given features x ∈ X, the HAI team makes decision y ∈ Y.
- We focus on *AI-assisted decision making*, where an AI model provides recommendation $y_m = m(x; \vartheta_m)$ to a human with their independent judgement $y_h = h(x; \vartheta_h)$, who makes final team decision $d = f(x, y_m, y_h)$.
- As an initial step to better factor human behavior, we propose to use a threshold-based team model, where humans utilizes AI only when their self confidence is sufficiently low (below ρ). *A higher ρ is associated with higher frequency for humans to rely on the AI model.*

$$f(\mathbf{x_i}, m(\mathbf{x_i};\theta_m), h(\mathbf{x_i};\theta_h)) = \begin{cases} h(\mathbf{x_i};\theta_h) & \text{if } \mathcal{C}_i > \rho \\ m(\mathbf{x_i};\theta_m) & \text{otherwise} \end{cases}$$

Training Complementary AI

- Standard training (1) optimizes AI model's independent performance, neglecting human's contribution to the decision making process, while complementary training (2) optimizes team performance.

$$\theta_m = \arg\min_{\theta'_m} \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x_i}, y_i) \in \mathcal{D}} \ell\left(m(\mathbf{x_i};\, \theta'_m), y_i\right) \qquad (1)$$

$$\theta_m = \arg\min_{\theta'_m} \mathcal{L}_{team} = \arg\min_{\theta'_m} \tfrac{1}{|\mathcal{D}|} \textstyle\sum_{(\mathbf{x_i}, y_i) \in \mathcal{D}} \ell\big(f(\mathbf{x_i}, m(\mathbf{x_i};\theta'_m), h(\mathbf{x_i};\theta_h)), y_i\big) \qquad (2)$$

- Under our assumed, threshold-based decision making model, we show that ***human-confidence-based instance weighting*** results in complementary training.

$$\theta_c = \arg\min_{\theta'_c} \sum_{(\mathbf{x_i}, y_i) \in \mathcal{D}} w_i \cdot \ell(m_c(\mathbf{x_i};\theta'_c), y_i)$$

- *If human decision maker is less confident about instance i than instance j, then instance i should be weighted at least as high as instance j (i.e, **$w_i \geq w_j$ if $C_i < C_j$**).*
- **Optimization for known self-confidence threshold:** *When the human decision maker uses a fixed and known self-confidence threshold ρ to determine the human-AI team joint decision, the team loss is minimized when **$w_i = 1[C_i \leq \rho]$**.*
- **Optimization for expected self-confidence thresholds:** *When the human decision maker draws a self-confidence threshold from a known distribution to determine the human-AI team joint decision, i.e., $\rho \sim f_T(\rho)$, the expected team loss is minimized when **$w_i = 1 - F_T(C_i)$**, where $F_T(\cdot)$ is the CDF for threshold ρ.*
  - Based on above, the heuristic method **$w_i = 1 - C_i$** turns out to be optimal when human decision maker draws self-confidence from a uniform distribution *$\rho \sim U[0,1]$*.

Evaluation

- Simulation studies on synthetic **College Admission** (whether to admit an applicant to college, with decision also influenced by group membership) and real-world **WoofNette** (5 easily recognizable objects and 5 difficult dog breeds from ImageNet) data.
- Persistent gains under varied self-confidence threshold distributions and human characteristics, including undesirable but common settings like ill-calibrated human self-confidence, making our solution particularly beneficial in more practical setups.