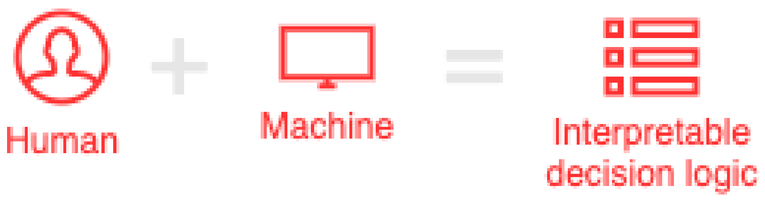
Co-creating a globally interpretable model with human input

Rahul Nair

IBM Research

Concept



Aim to learn an interpretable model for a supervised classification task jointly with humans. Aggregate **decision logic** rather than decision outcomes.

- *Complementarity:* Codify decision processes that may be difficult to learn from data along. Use broader context of the problem,
- *Interpretability:* Resulting model is still interpretable,
- *Coverage:* Capture situations not seen in training or factor in changes in regulatory requirements.

How?

Extend method proposed by Dash et al. (2018) on Boolean Rule via Column Generation (**BRCG**). **BRCG** is a mixed-integer program that seeks to find a subset of rules (w_k) that minimizes total Hamming loss, where the Hamming loss for each sample is the number of conjunctions that must be added or removed to classify it correctly, i.e.

$$\min_{\xi,w} \underbrace{\sum_{i\in P}\xi_i}_{\text{false negatives}} + \underbrace{\sum_{i\in Z}\sum_{k\in K_i} w_k}_{\text{false positives}}.$$

We consider two extensions to include human inputs by modifying the objective function with either user information in the form of complete conjunctions U

$$\min_{\xi,w} \underbrace{\sum_{i\in P}\xi_i + \sum_{i\in Z}\sum_{k\in K_i} w_k}_{\text{Machine objective (Hamming loss)}} + \underbrace{c_u n \sum_{k\in U}(1-w_k)}_{\text{Human inputs (violation penalty)}},$$

or with partial conjunction templates U' (with an associated notion of distance):

$$\min_{\xi,w} \sum_{i\in P}\xi_i + \sum_{i\in Z}\sum_{k\in K_i} w_k + \underbrace{c_p \sum_{k\in K} d(k,U')w_k}_{\text{distance penalty for partials}}.$$

Example: in a mortgage approval task, U can include a condition based on Loan-to-Value ratios of income, which is set by regulators:

$$(\text{LTV} \geq 90\%) \vee (\text{LoanAmount} \geq 3.5 \times \text{Income})$$

Acknowledgements

Thanks to Jonathan Epperlein and Anne-Marie Cromack for comments and review of an earlier draft of the paper. This work was partially funded by the European Union's Horizon Europe research and innovation programme under grant agreement no. 101070568 - AutoFair.

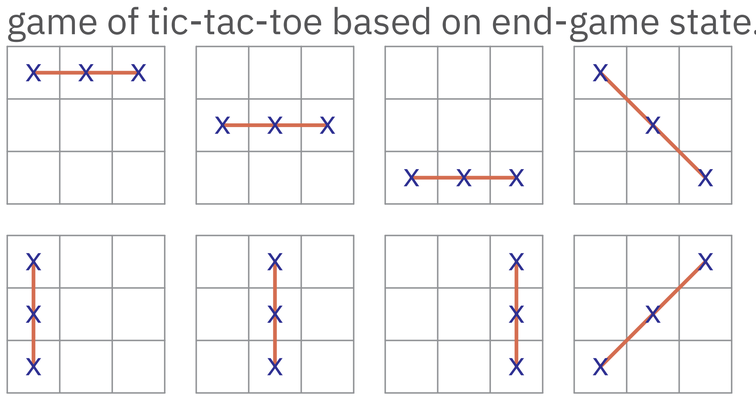
Evaluation

Resulting models are evaluated across two dimensions:

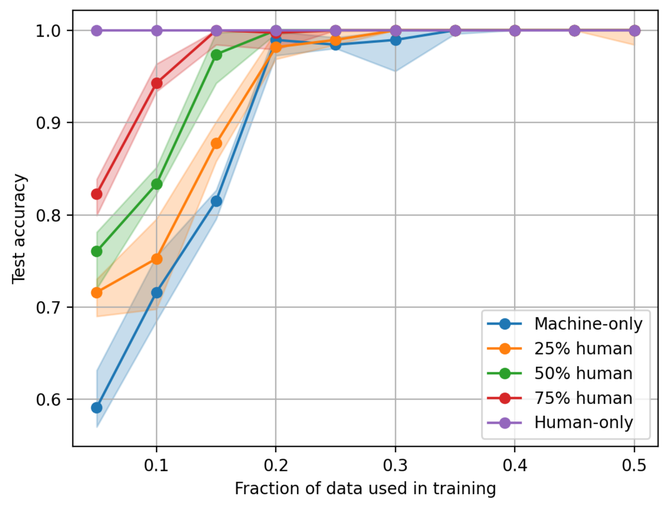
- *Performance:* Generalisation through test accuracy (only if test data exists to support it)
- *Interpretability:* Length and content of generated rules. Shorter rule sets are preferred and content is measured using semantic similarity to user provided inputs.

Example

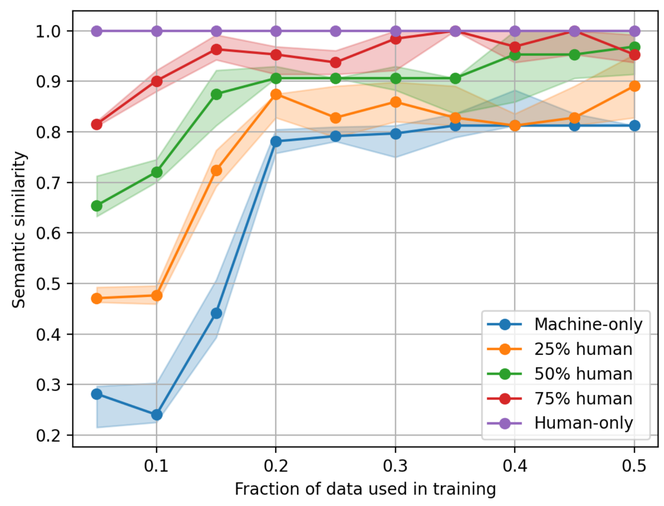
Noise-free classification task on if **x** wins a game of tic-tac-toe based on end-game state.



Vary the number of these rules known before rule induction and evaluate generalisation and interpretability.



(a) Median test accuracy



(b) Semantic similarity (higher is better)

Figure 1: Results for the noise-free classification task in tic-tac-toe
Empirical demonstration that human-assisted rule induction outperform machine only, for this noise-free task where rules are known perfectly.
Better rule semantics (e.g. diagonal rules often encoded by machine-only as absence of 'o', correct semantics induced with human inputs as presence of 'x').

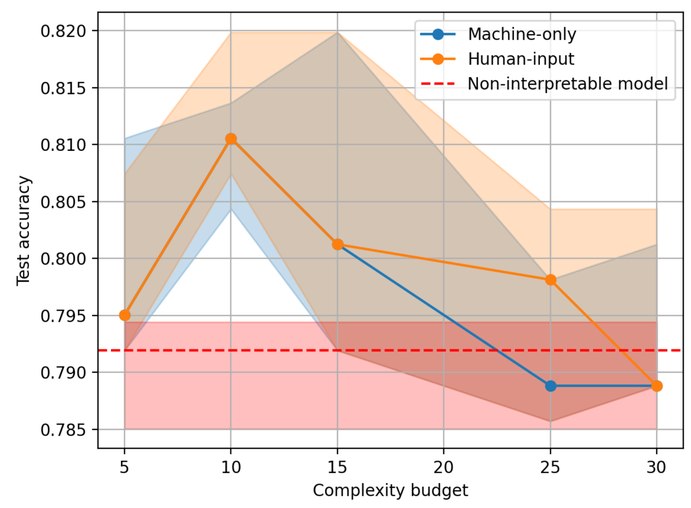
Application

Sepsis is a serious medical condition due to organ dysfunction brought about by infection. Rapid treatments are key to reduce risk.

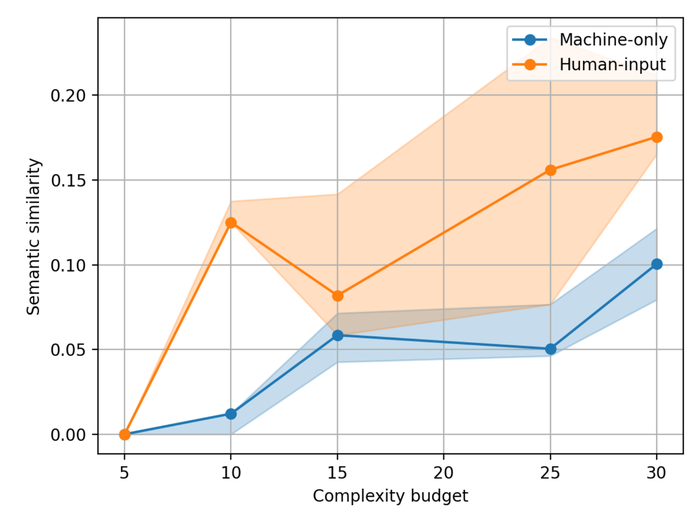
Human input For our experiments, we consider the conditions on administering drugs in a specific order, i.e.

$$\text{Survival Outcome} \leftarrow \text{Use Drug}_1 \wedge \text{Use Drug}_2$$

where, Drug_1 denotes set of drugs used for severe sepsis, and Drug_2 are drugs for mild sepsis. Drugs meant for severe sepsis cannot be administered before those used for mild sepsis, as this can lead to antibiotic resistance. Task is to predict survival condition of patients using 30-day mortality outcomes.



(a) Median test accuracy



(b) Semantic similarity (higher is better)

Figure 2: Results for the healthcare application of 30-day mortality of Sepsis patients
The human-assisted variant does no worse than the machine-only variant in terms of performance, but does better on rule semantics. Resulting model reflects domain-expert intuition and insight.

Limitations and Future Work

- Codifying human intuition is a challenge
- Disagreements between intuitions and data
- Lack of evidence hampers a proper evaluation

Paper