Interpretable Sequence Classification via Discrete Optimization

Maayan Shvo† Andrew C. Li Rodrigo Toro Icarte Sheila A. McIlraith†

Department of Computer Science, University of Toronto, Toronto, Canada

Vector Institute for Artificial Intelligence, Toronto, Canada

†Schwartz Reisman Institute for Technology and Society, Toronto, Canada

- Sequence classification is the task of predicting a class label given a sequence of observations.
- The class of problems we address are **symbolic** time-series classification problems that require discrimination of a set of potential classes.

Motivation:

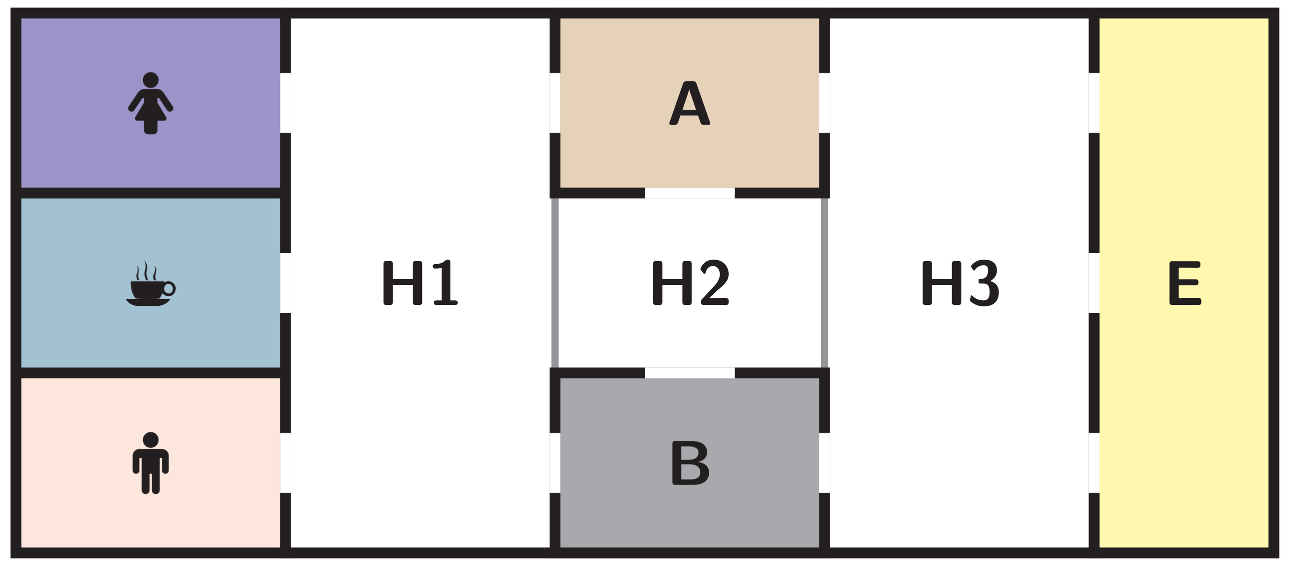
- Recurrent neural networks are state-of-the-art sequence classifiers, but the rationale for classification is difficult for a human to discern.
- In many applications (e.g. healthcare monitoring, malware detection), early classification is crucial to prompt intervention.

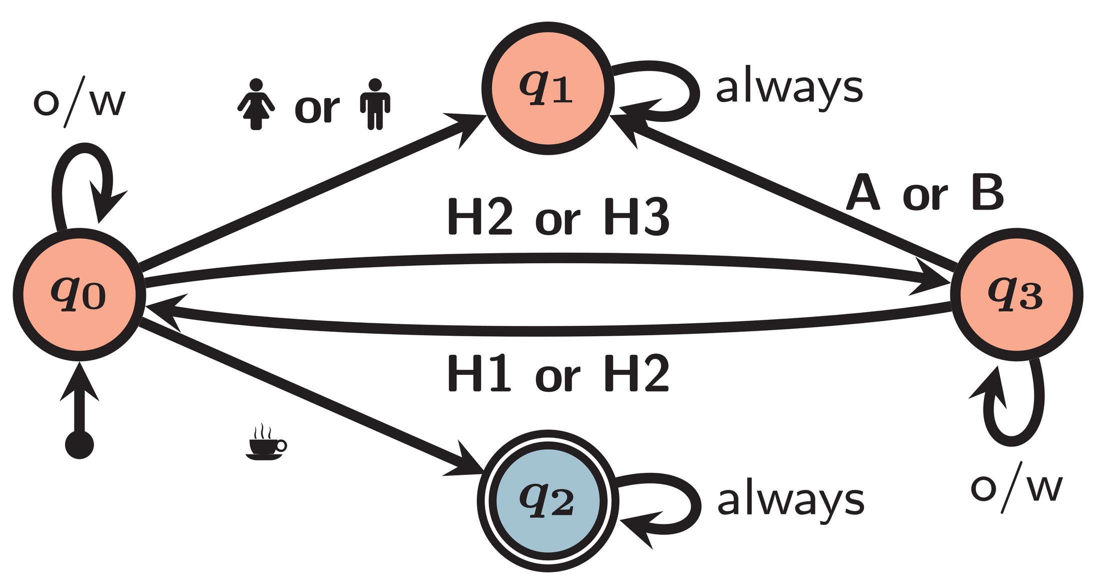
Contributions:

- We introduce **D**iscrete Optimization for **I**nterpretable **S**equence **C**lassification **(DISC)** which uses DFAs as **interpretable** sequence classifiers which favour **early classification**.
- We propose a novel discrete optimization approach for learning DFA classifiers that are **robust to noisy data** and are suitable for real-world domains.
- We outline how learned DFAs can support **explanation**, **counterfactual reasoning**, and **human-in-the-loop modification**.
- Experiments show that DISC achieves **comparable test performance to LSTM**, with the added advantage of being interpretable.

Example

Consider a goal recognition environment where the possible goals of the agent are going to an office (**A** *or* **B***), leaving the building (***E***), going to the restroom (🚺 or 🚹), or getting coffee (☕). The agent starts at* **A***, B, or* **E** *and takes the shortest (Manhattan distance) path to the goal.* **Right** *- a DFA classifier that detects whether or not the agent is trying to reach the goal ☕. A decision is provided after each new observation based on the current state:* **yes** *for the blue accepting state, and* **no** *for the red, non-accepting states.*





Learning DFAs for Sequence Classification

Learning one vs rest binary classifiers

- We train a separate DFA to recognize traces from each class.
- We specify a MILP model to find the DFA minimizing classification error.
- We regularize the number of non-self-loop transitions to prevent overfitting to handle noise.

Multiclass classification

- We perform Bayesian inference over the ensemble of DFAs (one per class).
- We infer a probability distribution over classes which is useful for confidence estimation.

Classifier Verification and Modification

- Temporal properties of the DFA classifier such as *"Neither 🚺 nor 🚹 occur before ☕"* can be straightforwardly specified in LTL and **verified** against the DFA using standard formal methods verification techniques.
- Our learned classifiers are also amenable to the inclusion of additional classification criteria, and the **modification** to the DFA classifier can be realized via a standard product computation.

Counterfactual Explanation

- In cases where a classifier does not return a positive classification for a trace, a useful explanation can take the form of a so-called *counterfactual explanation*.
- Given the trace $\tau =$ (**A, H2, H1,** 🚹), a possible counterfactual explanation is the edit operation (informally specified) REPLACE 🚹 WITH ☕ which transforms (**A, H2, H1,** 🚹) to (**A, H2, H1,** ☕). This explanation can then be transformed into a natural language sentence: *"The binary classifier would have accepted the trace had ☕ been observed instead of 🚹"*.

Experimental Evaluation - Datasets

Three goal recognition datasets:

- Crystal Island, a narrative-based game
- ALFRED, a virtual-home environment
- MIT Activity Recognition (MIT-AR)

Three behavior classification datasets:

- A dataset comprising replays of different types of scripted agents in the real-time strategy game StarCraft
- Two real-world malware datasets comprising 'actions' taken by different malware applications in response to various Android system events

Experimental Evaluation

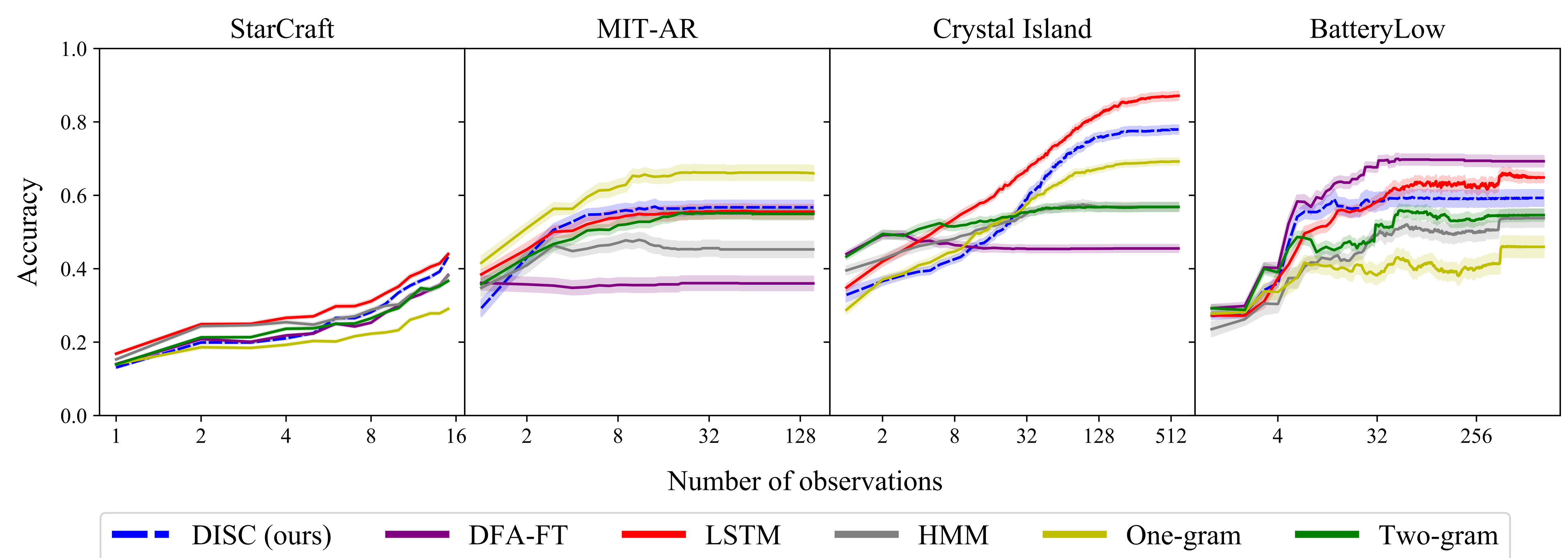


Figure: Test accuracy of DISC and all baselines as a function of earliness (number of observations seen so far) on one synthetic dataset (left) and three real-world datasets (three right). We report Cumulative Convergence Accuracy up to the maximum length of a trace. Error bars correspond to a 90% confidence interval.

Discussion

DFA-FT

- A DFA-learning approach that does not perform regularization, representative of existing work in learning DFAs.
- Qualitatively, the DFAs learned by DISC were orders of magnitude smaller than those learned by DFA-FT.
- DFA-FT often overfits to noise.

Pros

- **DISC often achieves near LSTM performance**, and outperforms other baselines.
- DISC learns interpretable models.

Cons

- DISC assumes the traces for each label can be recognized by a DFA (or equivalently, form a regular language) which does not always hold true.
- DISC cannot easily solve tasks requiring counting.