Domain-independent Plan Intervention

Sachini Weerawardhana, Darrell Whitley, Mark Roberts

AAAI/Plan, Activity and Intention Recognition (PAIR) Workshop

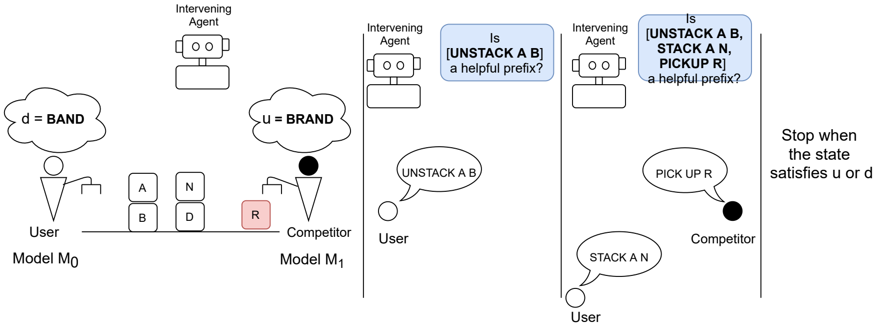
February, 9, 2021

The Main Takeaway

- The problem
  - An agent is executing a plan online (user)
  - The environment has some conditions that may allow the user's plan to be subverted (e.g., hidden information, an attacker)
  - A passive observer monitoring the agent(s) actions must recognize in advance the user's plan will have an undesirable outcome
- Research contribution
  - Use characteristics of the planning problem representation to learn when intervention is required

An Intervention Episode

Three agents: two actors (user, competitor) and one observer (intervening agent)



Research Question: How to identify the salient characteristics for deciding to intervene?

- Compare solutions that contain undesirable moves and solutions that do not.
- Two sources of extracting characteristics:
  - Intervention Graph
  - Plans sampled from the plan space

Intervention by Unhelpful Plan Prefix Recognition

Assumptions

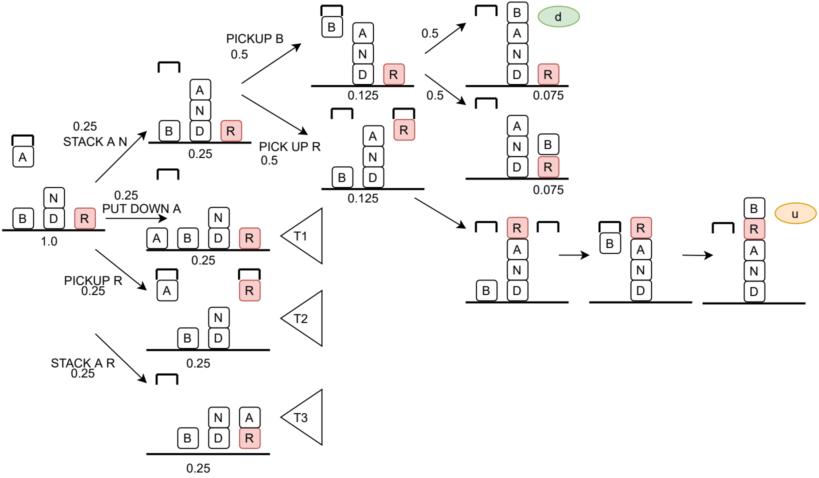
- d is user's goal, u is the undesirable goal
- The observer has full observability of the actors' actions.
- The observer knows about d and u and helps the user avoid u
- u is unknown to the user.
- d is unknown to the competitor (if present)
- User can not recognize effects of competitor's actions. Some of user's own actions may have hidden effects
- User follows a satisficing plan to reach d, and may reach u unwittingly

The recognition problem

Given the action history $(o_1, \ldots, o_{i-1})$, and the proposed action o_i must the prefix $(o_1, \ldots, o_{i-1})$ be flagged to help the user avoid u?

Unhelpful Plan Prefix Recognition - Intervention Graph

- Produce the intervention graph from current state (root) to u =(BRAND) and d =(BAND)



Compute features of critical actions (or sequences)

- **Risk** - Posterior probability of reaching u, when the user is trying to reach d
- **Desirability** - Posterior probability of reaching d, without passing u
- **Distance to d** - Mean number of edges between the root of the tree and d, which doesn't pass through u
- **Distance to u** - Mean number of edges between the root of the tree and u
- **Active attack landmarks%** - From the total number of predicates that must be true in any valid solution to the planning problem $\langle M, \mathrm{u} \rangle$, how many are true in the current state?

Unhelpful Plan Prefix Recognition - Sampling the Plan Space

- Instead of computing exact distances and probabilities compute an **estimated proximity to d and u**
- Use an automated planner to find two solution sets for $\langle M, \mathrm{d} \rangle$ and $\langle M, \mathrm{u} \rangle$
- Compute a *reference plan* $= \{$*observations* $+ \pi^*\}$,
  - π^* - cost optimal plan from the current state to the goal for d
- Compute the "distances" between the *reference plan* and the two solution sets
  - Is the reference plan more similar to sampled d plans or u plans?
  - plan distance metrics: action set distance, causal link distance, state sequence distance, edit distance.

Learning to Intervene

- Two feature sets:
  - intervention graph metrics
  - distance metrics from the sampled plans
- Use the feature sets to train classifiers to recognize unhelpful plan prefixes
  - Naive Bayes, K-nearest neighbor, Decision tree, Logistic regression
- Use the classifiers to recognize intervention in unseen problems

Classifier Performance

Reporting F-score $= \frac{TP}{TP+1/2(FP+FN)}$
Matthews Correlation Coefficient (MCC)

- Planning domains: BlocksWorld, EasyIPC, Ferry, Navigator, RushHour
- Classifiers with Intervention Graph features
  - High accuracy (F-Score, MCC $>$87%) for all domains
  - Uniform high accuracy across different classifiers
- Classifiers with plan distance metrics
  - Accuracy varies for different types of classifiers for each domain
  - Blocksworld problems - **K-nearest neighbor**
  - EasyIPC/Navigator problems -**Naive Bayes**
  - Rush Hour problems -**Decision Tree, K-nearest neighbor**
  - Lowest accuracy for Ferry domain problems

Intervention with Existing Goal Recognition Algorithms

- Use **plan recognition as planning** and **goal mirroring** to recognize likely goals given observations
- If the likely goal is u, then intervene
- Results
  - Reporting F-score and Matthews Correlation Coefficient (MCC) for benchmark planning domains (Blocks words, EasyIPC, Ferry)
  - Many false negatives/positives occur when the undesirable state is close to the desirable state in the state space.
  - Low accuracy in recognizing when intervention is required

Questions?

Sachini Weerawardhana
sachini@cs.colostate.edu

Computer Science Department
Colorado State University
Fort Collins, CO 80524, USA