Learning Structured Decision Problems with Unawareness

Craig Innes (craig.innes@ed.ac.uk),
Alex Lascarides (alex@inf.ed.ac.uk)

Institute for Language, Cognition and Computation
University of Edinburgh
\[ \mathcal{X} = \{ \text{Prec, Protein, Yield} \} \]

\[ \mathcal{A} = \{ \text{Grain, Fert} \} \]

\[ \text{scope}(\mathcal{R}) = \{ \text{Yield, Protein} \} \]

\[ \text{Pa}_{\text{Prot}} = \{ \text{Grain} \} \]

\[ P(\text{Prot} = p | \text{Grain} = g) = \theta_{p|g} \]
Why Unawareness?

\[ \mathcal{X}^0 \subseteq \mathcal{X}^+ \]
\[ \mathcal{A}^0 \subseteq \mathcal{A}^+ \]
\[ \text{scope}_0(\mathcal{R}) \subseteq \text{scope}_+^{+}(\mathcal{R}) \]
\[ \text{Pa}_{\text{Prot}} = \{ \text{Grain} \} \]
\[ P(\text{Prot} = p | \text{Grain} = g) = \theta_{p|g} \]
Contributions

Our agent learns an **interpretable model** of a decision problem **incrementally** via evidence from **domain trials** and **expert advice**.
Contributions

Our agent learns an **interpretable model** of a decision problem **incrementally** via evidence from **domain trials** and **expert advice**.

*Evidence may reveal actions/variables the agent was completely unaware of prior to learning.*
Types of Advice

1. **Advice on Better Actions**
2. Resolving Misunderstandings
3. Unexpected Rewards
4. Unknown Effects
Contextual Advice - Better Action

If agent’s performance in last $k$ trials is below threshold $\beta$ of true policy $\pi_+$, say:

- $a' = \langle A_1 = 0, A_2 = 1, A_3 = 0 \rangle$ rather than $a_t$

- $A_3$ is part of the problem ($A_3 \in A$)

- $A_3$ is relevant ($\exists X \in \text{scope}(R), \text{anc}(A_3, X)$)

- There exists a better reward ($\exists s, s[B_t] = s[t[B_t] \land R_+(s)] > r_t$)

- $a'$ has a greater expected utility than $a_t$ ($\text{EU}(a')|s) > \text{EU}(a_t|s)$)
If agent’s performance in last $k$ trials is **below threshold** $\beta$ of true policy $\pi_+$, say:

“At time $t$ you should have done $a' = \langle A_1 = 0, A_2 = 1, A_3 = 0 \rangle$ rather than $a_t$”
If agent’s performance in last $k$ trials is **below threshold** $\beta$ of true policy $\pi_+$, say:

“At time $t$ you should have done $a' = \langle A_1 = 0, A_2 = 1, A_3 = 0 \rangle$ rather than $a_t$”

- Action variable $A_3$ is part of the problem ($A_3 \in \mathcal{A}$)
- $A_3$ is relevant ($\exists X \in \text{scope}(\mathcal{R}), \text{anc}(A_3, X)$)
- There exists a better reward ($\exists s, s[\mathcal{B}^t] = s_t[\mathcal{B}^t] \land \mathcal{R}_+(s) > r_t$)
- $a'$ has a greater expected utility than $a_t$ ($EU(a'|s) > EU(a_t|s)$)
Conserving Previous Beliefs

\[ P(\text{Pa}_{\text{Yield}} \mid D_{0:t}) \]

\[ \text{Pa}_{\text{Yield}} = \emptyset \]

\[ \text{Pa}_{\text{Yield}} = \{ \text{Fert} \} \]

\[ \vdots \]

\[ \text{Pa}_{\text{Yield}} = \{ \text{Fert, Prec, Grain} \} \]
Conserving Previous Beliefs

\[
P(P_{\text{Yield}}|D_{0:t})
\]

\[
P_{\text{Yield}} = \emptyset
\]

\[
P_{\text{Yield}} = \{ \text{Fungus} \}
\]

\[
P_{\text{Yield}} = \{ \text{Fert} \}
\]

\[
P_{\text{Yield}} = \{ \text{Fert, Fungus} \}
\]

\[
\vdots
\]

\[
P_{\text{Yield}} = \{ \text{Fert, Prec, Grain} \}
\]

\[
P_{\text{Yield}} = \{ \text{Fert, Prec, Grain, Fungus} \}
\]
Conserving Previous Beliefs

\[ P(\text{Pa}_{\text{Yield}}|D_{0:t}) \]

\[ \text{Pa}_{\text{Yield}} = \emptyset \]

\[ \text{Pa}_{\text{Yield}} = \{\text{Fungus}\} \]

\[ \text{Pa}_{\text{Yield}} = \{\text{Fert}\} \]

\[ \text{Pa}_{\text{Yield}} = \{\text{Fert, Fungus}\} \]

\[ \vdots \]

\[ \text{Pa}_{\text{Yield}} = \{\text{Fert, Prec, Grain}\} \]

\[ \text{Pa}_{\text{Yield}} = \{\text{Fert, Prec, Grain, Fungus}\} \]

\[ P_{\text{new}}(\text{Pa}_X) = \begin{cases} 
(1 - \rho)P_{\text{old}}(\text{Pa}_X|D_{0:t}) & \text{if Fungus} \notin \text{Pa}_X \\
\rho P_{\text{old}}(\text{Pa}'_X|D_{0:t}) & \text{if } \text{Pa}_X = \text{Pa}'_X \cup \{\text{Fungus}\} 
\end{cases} \]
Experiments

Randomly Generated Networks: 12 - 36 Variables

- 12 - 36 Variables
- 3000 Trials
- $\epsilon$-greedy strategy
- Expert Aid $\beta = 0.1$
Results

![Graph showing cumulative reward over time for different policies: default, truePolicy, and random.](image)

The graph illustrates the cumulative reward for different policies over time. The x-axis represents time (t) ranging from 0 to 3000, and the y-axis represents the cumulative reward, which ranges from 0 to 60. The lines indicate the performance of the default, truePolicy, and random policies, with the default policy showing the highest cumulative reward over time.
Results

The chart shows the cumulative reward over time for different tolerance settings. The x-axis represents time (t) ranging from 0 to 3000, and the y-axis represents the cumulative reward. There are three lines in the chart:

- Blue line labeled "default".
- Orange line labeled "lowTolerance".
- Green line labeled "highTolerance".

The charts indicate that the "default" setting consistently achieves the highest cumulative reward, followed by "lowTolerance" and then "highTolerance".
Paper

Learning Structured Decision Problems with Unawareness

Authors

Craig Innes (craig.innes@ed.ac.uk)
Alex Lascarides (alex@inf.ed.ac.uk)

Poster Session:

6:30pm-9pm, Pacific Ballroom #35