Explaining Automated Policies for Sequential Decision Making

February 16, 2010
University of Kentucky, Lexington

Presented by Pascal Poupart
University of Waterloo, Canada

Joint work with Omar Zia Khan and Jay Black
Sequential decision making

- Fault diagnosis, inventory management (OR)
- Medical diagnosis (health informatics)
- Course selection advising (recommender systems)
- Robotic control
- Web optimization

- Difficult to optimize policy
  - Uncertain action effects
  - Multiple/complex objectives
  - Repeated/sequential decision points
Automated Policy Generation

- **Solution:**
  - Harness the power of machines
  - Automated policy optimization

- **Problem:**
  - How can we ensure user trust?
  - How can we verify the correctness of the model and resulting policy?

- **Contribution: policy explanation**
  - Generic approach to explain the choice of action
Outline

- Background
- Automated policy explanation
- Experiments and Sample Explanations
- User Study
- Conclusion and Future Work
Markov Decision Processes

• General framework for sequential decision making
• Formalized in Operations Research in the 1950s
• Automated policy optimization
• Today: one of the most popular approaches
• But, no generic technique to explain resulting policies

• This talk: automated explanation of MDP policies
  – Generic, problem independent technique
  – Minimal and sufficient explanations
**MDP Graphical Representation**

Transition function: \( \text{Pr}(s_t|s_{t-1},a_{t-1}) \)

Reward function: \( R(s_t,a_t) \)

Solution: policy \( \pi \) maximizes expected total rewards
Policy Evaluation

• Optimal policy maximizes expected rewards

\[ V^*(s) = \max_a \left[ R(s, a) + \gamma \sum_{s' \in S} \Pr(s'| s, a) V^*(s') \right] \]

• Occupancy frequencies
  – Expected number of times a state is visited by executing a policy given a starting state

\[ \lambda_{s_0}^\pi (s') = \delta(s', s_0) + \gamma \sum_{s \in S} \Pr(s'| s, \pi(s)) \lambda_{s_0}^\pi (s) \]

• Value of policy can be computed using terms that are products of occupancy frequencies and rewards

\[ V_{s_0}^\pi (s) = \sum_{s \in S} \lambda_{s_0}^\pi (s) R(s, \pi(s)) \]
Difficulty in Explanation

- Policy computed using complex numerical techniques
- Most primitive explanation
  - Action maximizes expected utility

- Issues
  - Many factors contribute to utility
  - Numerical value of utility only reflects preference (unless it represents something tangible like money or time)
  - Computation of expected utility is complex
Overview of Approach

• Use pre-defined templates populated at run-time
  – Not concerned with natural language generation
  – Number of templates identified such that explanations are sufficient yet minimal

• Report occupancy frequency of certain states
  – Focus on states with high/low reward
  – But do not report numerical utilities (harder to grasp)
Templates

• T1: Action *actionName* is the only action that is likely to take you to *var1=val1, var2=val2, var3=val3* about \( x \) times which is higher (or lower) than any other action

• T2: Action *actionName* is likely to take you to *var1=val1, var2=val2, var3=val3* about \( x \) times which is as high (or low) as any other action

• T3: Action *actionName* is likely to take you to *var1=val1, var2=val2, var3=val3* about \( x \) times
Minimal Sufficient Explanations

- Multiple templates possible for non-optimal actions
  - Non-optimal action may have highest frequency of reaching a state/scenario
  - May not guarantee highest expected utility

- Explanation with single template may be insufficient
- Explanation with all templates may be overwhelming

- Identify optimal number of templates to create a “Minimal Sufficient Explanation”
Minimal Sufficient Explanations

• Utility of explanation

\[ V_{Explanation} = \sum_{i} r(s_i) \lambda_{s_0}^{\pi^*}(s_i) + \sum_{j} r_{\min} \lambda_{s_0}^{\pi^*}(s_j) \]

\[ \text{templates} \quad \text{no template} \]

• Minimal sufficient explanation
  – Fewest templates with utility greater than any other action choice

\[ V^{\pi^*} \geq V_{MSE} > V^{\pi'} \]
MSEs for Factored MDPs

- State space defined by a set of variables
  - Scenarios defined as set of states resulting from assigning values to a subset of variables
- Reward function can also be decomposed
- Value of policy can be computed using scenarios
  \[ V^\pi(s) = \sum_k \sum_{r \in \text{dom}(R_k)} r \lambda^\pi_{s_0}(sc_{R_k=r}) \]
- Value of explanation can also be computed using scenarios
  \[ V_{\text{Explanation}} = \sum_i r_i \lambda^\pi_{s_0}(sc_i) + \sum_j r_{\text{min}} \lambda^\pi_{s_0}(sc_j) \]
Numerical Example

- **Rewards**
  - \( R(\text{Courses}=6) = 100, R(\text{Courses} \neq 6) = 0 \)
  - \( R(\text{Areas}=3) = 100, R(\text{Areas} \neq 3) = 0 \)

- **Optimal action**
  - \( \lambda(\text{Courses}=6) = 0.67, \lambda(\text{Courses} \neq 6) = 0.33 \)
  - \( \lambda(\text{Areas}=3) = 0.95, \lambda(\text{Areas} \neq 3) = 0.05 \)
  - \( V^* = 100 \times 0.67 + 0 \times 0.33 + 100 \times 0.95 + 0 \times 0.05 = 162 \)

- **2\text{nd} best action**
  - \( \lambda(\text{Courses}=6) = 0.25, \lambda(\text{Courses} \neq 6) = 0.75 \)
  - \( \lambda(\text{Areas}=3) = 0.68, \lambda(\text{Areas} \neq 3) = 0.32 \)
  - \( V^\text{2nd} = 100 \times 0.25 + 0 \times 0.75 + 100 \times 0.68 + 0 \times 0.32 = 93 \)

- **Minimal sufficient explanation**
  - \( V_{\text{MSE}} = (100 \times 0.95) + (0 \times 0.67 + 0 \times 0.33 + 0 \times 0.05) = 95 \)
Algorithm

1. For each R do 
   a. For each \( r \in \text{dom}(R) \) do 
      i. Compute occupancy frequency: \( \lambda(\text{sc}_{R=r}) \) 
      ii. Template value: \( r \lambda(\text{sc}_{R=r}) \) 
2. Order templates in decreasing value 
3. Show minimal # of templates to ensure sufficient explanation
Invariance of MSEs

Proposition: MSEs remain invariant under affine transformations of reward function

Proof:
- Occupancy frequencies add up to horizon $h$
- Substitute $r$ with $r+c$

\[
\hat{V}^{\pi}(s) = \sum_k \sum_{r \in \text{dom}(R_k)} (r + c) \lambda_{s_0}^{\pi}(sC_{R_k} = r)
\]

\[
= V^{\pi}(s) + c \sum_k \sum_{r \in \text{dom}(R_k)} \lambda_{s_0}^{\pi}(sC_{R_k} = r)
\]

\[
= V^{\pi}(s) + cKh
\]
Experimental Setup

• Course Advising MDP
  – Choose best combination of courses
  – 117.4 million states with 21 actions
  – Transition model generated from historical data
  – Reward for different requirements of degree
  – Horizon is 3, no discounting

• Handwashing MDP (adapted from Hoey et al 2007)
  – Assist people with dementia in handwashing
  – 207,360 states, 25 actions
  – Horizon is 100, and discount factor is 0.95
Sample Explanations

• Action TakeCS343&CS448 is the best because:
  – It is likely to take you to CoursesCompleted=6, TermNumber=Final, about 0.86 times which is as high as any other action

• Action DoNothingNow is the best because:
  – It is likely to take you to handswashed=yes, planstep=Clean&Dry, about 0.71 times which is higher than any other action
  – It is likely to take you to prompt=NoPrompt about 12.71 times which is as high as any other action
Experimental Results

Course Advising Domain (Max Terms =4, Experiments=182)

<table>
<thead>
<tr>
<th>Terms in MSE</th>
<th>1</th>
<th>2</th>
<th>3–4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>134</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>Mean (STD)</td>
<td>0.46 (0.41)</td>
<td>0.81 (0.24)</td>
<td>-</td>
</tr>
</tbody>
</table>

Handwashing Domain (Max Terms =19, Experiments=382)

<table>
<thead>
<tr>
<th>Terms in MSE</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7–19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0</td>
<td>142</td>
<td>94</td>
<td>119</td>
<td>2</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Mean (STD)</td>
<td>-</td>
<td>0.51 (0.22)</td>
<td>0.62 (0.10)</td>
<td>0.68 (0.04)</td>
<td>0.61 (0.15)</td>
<td>0.69 (0.05)</td>
<td>-</td>
</tr>
</tbody>
</table>
User Study

- Recruited volunteers to evaluate automatically generated explanations for course advising MDP
- Objective
  - Evaluate our explanations
  - Compare with advisor explanations
- Demographics
  - 37 undergrad and grad students participated from CS
  - 5 explanations shown to each student
    - 3 generated using our technique
    - 2 similar to those offered by human advisors
## User Study Setup – Existing State

<table>
<thead>
<tr>
<th>Book-keeping Information</th>
<th>Core</th>
<th>Electives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term Number</td>
<td>4A</td>
<td>cs246 = Good</td>
</tr>
<tr>
<td>CGPA</td>
<td>Good</td>
<td>cs251 = Good</td>
</tr>
<tr>
<td>Systems/SE Area Covered</td>
<td>Yes</td>
<td>cs341 = Average</td>
</tr>
<tr>
<td>Applications Area Covered</td>
<td>No</td>
<td>cs350 = Good</td>
</tr>
<tr>
<td>Math Area Covered</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Electives Completed</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
User Study Setup – Sample Explanation

Taking \textit{cs348 \& cs370} is the best action because:

- You would be in a state with \textit{Electives Completed}=6 by end of \textit{Term Number}=4B with about \textit{79\%} chance which is as high as any other combination of courses.

- You would be in a state with \textit{Systems/SE Area Covered}=Yes, \textit{Mathematics Area Covered}=Yes, \textit{Applications Area Covered}=Yes, by the end of \textit{Term Number}=4B with about \textit{74\%} chance.
Effectiveness of MSEs

Total Responses 37×3 = 111

- Easy to Understand: 65 Yes, 46 No
- Provides Extra Info.: 79 Yes, 32 No
- Seems Accurate: 84 Yes, 27 No
- More Info. Needed: 77 Yes, 34 No
Comparison with Human Advisor Explanations

Total Responses = 37

- More Convincing
  - MDP-Based Explanation: 4
  - Advisor's Explanation: 12
  - Both Explanations: 21

- More Trust-worthy
  - MDP-Based Explanation: 6
  - Advisor's Explanation: 17
  - Both Explanations: 14

- Easier to Understand
  - MDP-Based Explanation: 1
  - Advisor's Explanation: 24
  - Both Explanations: 12
User Study – Results

• MSEs provide extra information and are trustworthy

• Human advisor explanations easier to understand

• Combination of MSE with human advisor is most preferred option

• Useful as a planning tool for students
Conclusion

• Domain-independent explanations for recommendations from MDP policies
  – Generated by populating pre-defined templates
  – No reference to numerical value of utility
  – Computed minimal set of explanations that completely justify the recommendation

• No additional effort needed from MDP designer
• User study indicates benefits of explanations

Future Work

• Inject domain-specific information in explanations
  – Represent domain-specific information in a domain-independent manner

• Explain effect of discount factor in explanations

• Extend explanations to POMDPs
  – Cater for observation function and distribution over initial state instead of single starting state
My Research Interests

• Areas
  – Reasoning under uncertainty
    • Sequential decision making (MDPs, POMDPs)
  – Machine learning, vision, NLP

• Application domains
  – Health informatics
    • Smart walker project
    • Symptom monitoring for Alzheimer’s disease
  – Document clustering
    • Unsupervised cluster labelling
Graduate Studies at U of Waterloo

• CS endowment of $25 million
  – Donor: David Cheriton (Waterloo PhD, Stanford prof.)
  – $1 million/yr for research & graduate studies

• CS is in the Faculty of Math
  – In AI, statistics and optimization are key
  – Easy interaction with dept. of Statistics and Combinatorics & Optimization.

• Start your own company
  – IP belongs to the creators (not the University)
  – Spinoffs: RIM, Maple, Open Text, etc.
  – Technopark on campus